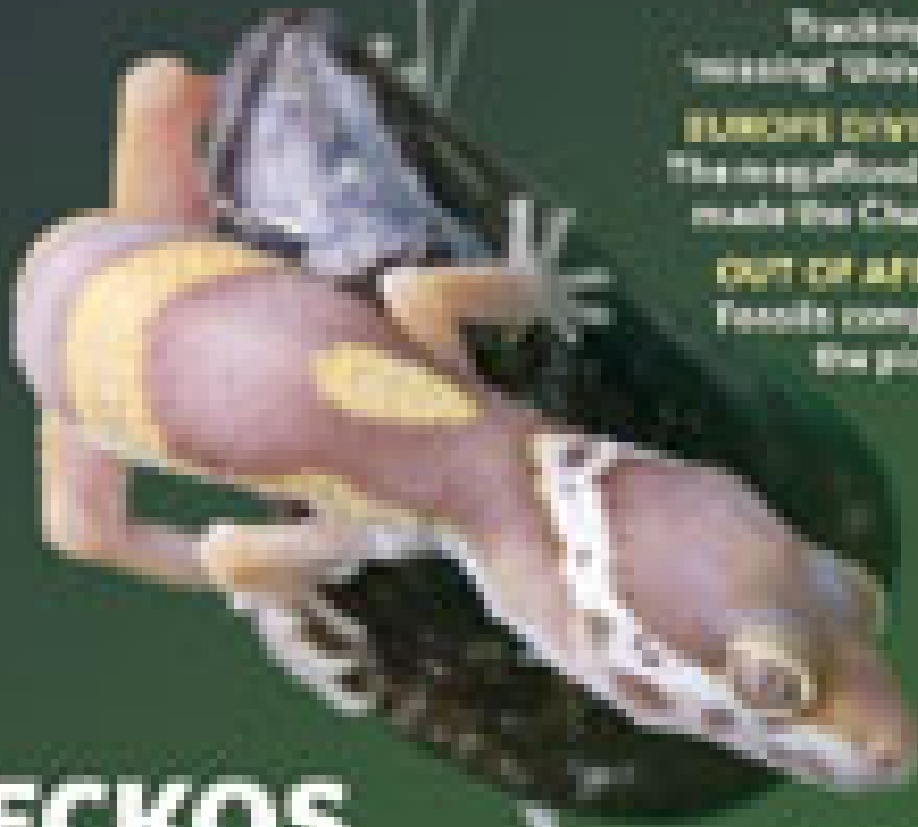
nature



THE DARK SIDE
Tracking the 'missing' Universe

EUROPE DIVIDED
The mega-flood that made the Channel

OUT OF AFRICA
Fossils complete the picture

GECKOS WITH MUSSEL

Combining to produce a powerful new biomimetic adhesive

NATURE JOBS
Research highlights

nature

www.nature.com/nature

Vol 448 | Issue no. 7151 | 19 July 2007

A unifying force

The questions to be explored at the Large Hadron Collider offer a chance to rekindle public interest in the fundamental principles of the Universe in which we live.

The Large Hadron Collider (LHC) that is nearing completion outside Geneva, Switzerland, is a testament to two of the greatest human qualities: a fascination with the workings of the Universe and an ability to cooperate to achieve shared goals. After the machine — now the centrepiece of Europe's particle-physics laboratory, CERN — goes into operation next year, it should allow experimentalists for the first time in decades to blaze paths into areas on which settled theory stands silent (see page 269).

The physicists exploring this new world will do so as a unified global community. Although CERN is broadly a European achievement, in which the continent can take great pride, the LHC and its attendant detectors have received contributions in cash or kind from more or less every country with the capability to participate in this sort of frontier science. The LHC therefore sets a new high-water mark in disinterested global cooperation.

But it is important to distinguish disinterested from uninterested. The nations of the world are not investing in the LHC because they necessarily expect to see returns similar to those physics delivered in the twentieth century (most obviously and terribly in the realm of nuclear weapons). At the same time, it is becoming apparent that such investments in physics no longer elicit quite the thrill that they once did among the general public.

Interest in the advances made at the LHC's immediate predecessors — the Large Electron Positron Collider at CERN and the Tevatron at America's Fermilab — drops off fairly steeply as one leaves the precincts of high-energy physics. Although the public likes the idea that scientists are making fundamental progress, the advances made in particle physics can seem increasingly far removed from ideas that resonate in the common imagination. In the first half of the twentieth century, the nucleus, relativity, the quantum and the uncertainty principle were quickly imbued with cultural meaning far beyond their scientific context. Gauge symmetries and the Higgs boson have yet to acquire such broader, symbolic importance.

Many physicists will consider it fanciful to suggest that they should — not least because the meanings that became associated with such concepts outside the realms of physics have often been far removed from, or in direct contradiction of, their scientific meaning.

At the same time, it is hard to ask people to spend the large sums needed to explore the frontiers of particle physics if they do not have some sense of investment in the questions that it asks. Here, the LHC offers an opportunity to re-establish a resonance between particle physics and the broader culture in which it sits. There is a strong case that the Universe is made up in large part of 'dark' matter and energy, quite unlike constituents that we can observe directly (see pages 240 and 245). It is possible that observations made by the LHC's detectors will speak directly to the nature of these, perhaps even producing in the laboratory some of the dark matter that seems to dominate the spiralling and clustering of galaxies.

"Advances made in particle physics can seem increasingly far removed from ideas that resonate in the common imagination."

The discovery of the hidden constituents of the Universe is a grand task, and has an imaginative appeal worthy of the global effort under way at CERN. And an ever stronger bond between the study of fundamental forces and particles on the one hand, and the structure and history of the Universe on the other, bodes well for the future of particle physics in other ways, too. Magnificent though machines such as the LHC are, they will necessarily be few and far between. It is important to have alternative ways forward that are less resource-intensive. Particle-flavoured astrophysics offers many opportunities along these lines, as do small-scale experiments looking for phenomena such as neutrinoless double-beta decay (see page 232).

CERN's grandeur comes in part from the paired missions of unification that it is embarked on — a theoretical unification of the phenomena of physics and a practical unification of the scientific aspirations of the world. Unifying the very large with the very small adds to the excitement generated, and should unite the imaginations of the world, helping to restore the prized position our culture has reserved for fundamental physics. ■

Transmission lines

Field trials of AIDS prevention methods are as essential as they are politically awkward.

More people than ever have access to effective AIDS treatments. But the virus will never be contained without more effective measures to prevent transmission — and the need for measures that can be initiated by women is especially urgent.

Unfortunately, the run-up to this year's International AIDS Society (IAS) conference in Sydney, Australia, has been dominated by negative research results concerning female-initiated prevention. But scientists and advocates should keep working resolutely together to make sure that testing of such methods continues apace.

The vast majority of new HIV infections in Africa, where the pandemic is most severe, occur through heterosexual transmission. But women are often powerless to negotiate the use of condoms — by far the best way to prevent infection. This is the impetus behind the clinical trials now testing alternative female-initiated prevention techniques. These include microbicides — gels or creams applied to the vagina to block infection; barrier methods, such as diaphragms;

and methods that can be used by both men and women, including preventative drugs.

It has been a long, difficult slog to get any of these methods into effective field trials, making the recent negative results doubly disappointing. In January, two trials of the microbicide cellulose sulphate were stopped when an interim analysis suggested that the product might make women more vulnerable to HIV. The product was the third microbicide to fail in efficacy studies and the second that seemed to increase the risk of HIV. And on 12 July, a team of researchers in South Africa, the United States and Zimbabwe reported that latex diaphragms used with condoms did not protect more women from HIV than condom use alone. On 25 July, investigators of the failed cellulose sulphate trials are expected to unveil their final data analysis at the IAS meeting — a step that will very probably spell the end for that particular product.

Looking forward, there is tension in the field over how best to conduct the next microbicide trials (see *Nature* **448,** 110–111; 2007). The danger is that further bad news will see funders lose their appetite for research on female-initiated prevention methods, so there is tremendous pressure to avoid more failures. This field has always been a difficult sell for policy-makers in any case: as long-time advocate Lori Heise of the Global Campaign for Microbicides says, it's about "women, vaginas and sexuality" — not topics that government officials especially want to air in public.

But developing and testing such measures will take a long time. There is no HIV vaccine in sight, either, but researchers seldom consider abandoning the quest for one. Product development is even more difficult than usual for female-initiated prevention methods, because testing them requires dealing with issues related to intimacy, cultural expectations and interpersonal relationships. It is hard for researchers to navigate these types of issues. Some see a more thorough investigation of all the circumstances surrounding a proposed intervention as a way around this. In a declaration circulating ahead of the Sydney meeting, which begins on 22 July, hundreds of scientists are calling for 10% of all HIV programme funds to be dedicated to such approaches.

"Women, vaginas and sexuality are not topics that government officials want to air in public."

But there is already a paucity of funding for proven prevention methods, according to a June report by the Global HIV Prevention Working Group. And a study released last week found that large-scale prevention programmes are the most cost-effective (E. Marseille *et al. BMC Health Services Res.* **7,** 108; 2007). It is clear that more resources should be directed at delivering the methods that work and at improving communication with the communities involved, to ensure both that existing prevention methods are used and that future trials will be conducted in optimal circumstances.

Dedicated researchers already know this. The principal investigators for a trial of a new microbicide gel containing the antiretroviral drug tenofovir, for instance, had extensive discussions with women before setting the dosing schedule for their drug. Such preparation is just as important as continued support for the search for good, female-initiated HIV prevention methods. With dedicated work on both fronts, researchers and advocates can be confident of finding the solutions that will control the pandemic and help women stay healthy. ■

Dedicated to science

Hands off the Commons Select Committee on Science and Technology.

What's in a name? That's one of the questions political leaders have to consider when they allocate titles to, and divisions between, government departments. The process is echoed when parliaments or other representative bodies set up committees to keep an eye on the activities of those departments.

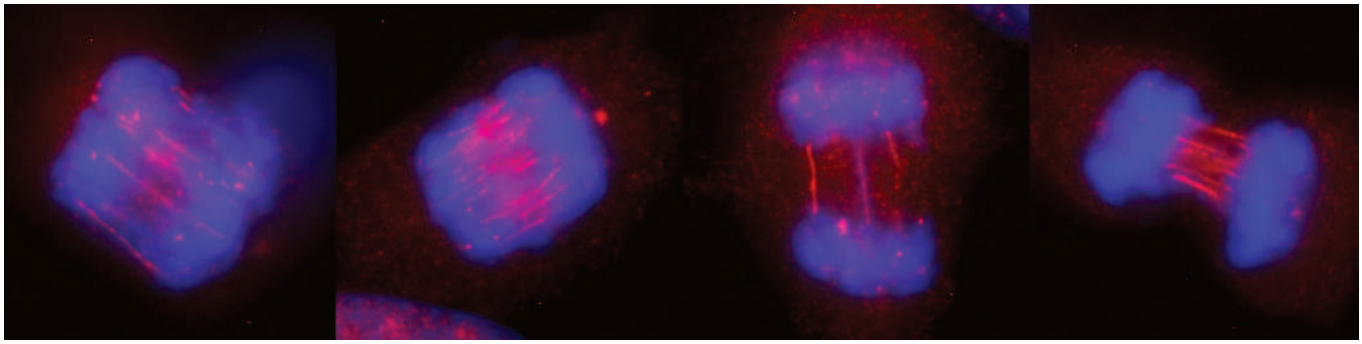
Every nation has its own approach to this, and some parliaments, including those of France and Germany, struggle to exercise much oversight at all. The UK House of Commons and the US House of Representatives have each, in very different circumstances, evolved committees that look expressly at science and technology questions. These committees perform a valuable role. By virtue of their very names, as well as their briefs, their remit centres on scientific and technological facts. Their staff and their members tend, on the whole, to be interested in such facts. These days, with the 'reality-based community' under steady attack from those who prefer to base their positions on dogma rather than on hard information, that's a rare blessing.

It is true that other committees, which sometimes have more heft, also consider issues related to science and technology. The Committee on Energy and Commerce in the House of Representatives and the Defence Committee at Westminster, for example, are both highly influential. There is an argument that discourse on scientific questions is best conducted where it matters most. But the reality in these more heavily politicized surroundings is that such discourse often does not take place at all.

It has been reported that the British government would like to wind up the Select Committee on Science and Technology and place its responsibilities in a new committee with a wider remit, dealing also with education and innovation, in line with an ongoing reorganization of the government's own departments (see page 236). This plan is doubly troubling because in Britain, where parliamentary committees are young and not especially powerful, the executive branch of the government can dictate the committee portfolios. Gordon Brown, the new prime minister, can informally tell the Leader of the House what is to be done, and it will happen.

It just seems too convenient that the science and technology select committee sometimes sheds light on inconvenient truths (such as the technical feasibility of politically attractive schemes for identity cards). There is no requirement — procedural, constitutional or in terms of precedent — for select committees to map directly on to particular government departments. If Brown wants intelligent and proactive oversight by parliament, as he has professed to do in his first few days in office, he should leave the House of Commons Select Committee on Science and Technology well alone. ■

RESEARCH HIGHLIGHTS



K.-L. CHAN & I. D. HICKSON

A bridge too far

EMBO J. doi:10.1038/sj.emboj.7601777 (2007)
Newly copied chromosomes stay entangled for longer during cell division than once thought, researchers report.

Ian Hickson and his team at the Weatherall Institute of Molecular Medicine in Oxford, UK, studied cells from patients with an inherited cancer predisposition called Bloom's syndrome. This arises from faults in a protein called BLM.

Duplicated chromosomes are pulled apart during the last stage of cell division, known as anaphase. In cells with the faulty protein, the copied chromosomes remain tethered together with large 'bridges' of DNA for longer than in normal cells. The researchers suggest that snapping of these bridges may cause the chromosome instability seen in Bloom's patients.

They also showed that BLM protein is associated with very fine DNA bridges between chromosomes in normal cells (shown pink in the image above), and speculate that the healthy protein helps untangle duplicated chromosomes.

BIOMECHANICS

Top legs

Phys. Rev. E **76,** 017301 (2007)
What's as sticky as a gecko and walks on water better than a water strider? According to Cheng Wei Wu of Dalian University of Technology in China and co-workers, the answer is a mosquito.

These researchers studied mosquitoes under the electron microscope. They saw legs covered with ribbed scales that confer high water repellence, and feet covered with hundreds of hair-like setae, which, like those of a gecko, penetrate into microscopic cracks to grip the smoothest of surfaces. The nanostructured legs are so efficient at supporting mosquitoes on water that they can hold up to 23 times the mosquito's body weight, whereas water striders can support only 15 times their own weight.
(For more about geckos, see p. 338.)

MICROBIOLOGY

Fungal attack

Proc. Natl Acad. Sci. USA **104,** 11772–11777 (2007)
Reactive oxygen species (ROS) defend plants against pathogens, but they may also be crucial for plant infection.

Nicholas J. Talbot of the University of Exeter, UK, and his colleagues found that *Magnaporthe grisea*, the fungus responsible for rice blast disease, requires two ROS-generating enzymes before it can damage a plant. Both enzymes are NADPH oxidases (Nox).

Although the fungus produces ROS in multiple ways, ROS generation through Nox was found to be important for the formation of specialized infection structures known as appressoria. Fungi with mutations that inactivate the two Nox develop appressoria that are not functional and therefore do not cause plant disease. ROS may aid infection by mediating changes in cell-wall biochemistry.

PLANETARY SCIENCE

Moon dust

Geophys. Res. Lett. **34,** L13203 (2007)
Thirty-five years after the Apollo 17 astronauts left the Moon (pictured below), researchers have mapped some aspects of the geology of the craft's landing site using the Hubble Space Telescope. The feat demonstrates a method that may help to unravel the Moon's volcanic history.

NASA-JPL

Some volcanic basalts are rich in the mineral titanium dioxide (TiO_2). Mark Robinson of Arizona State University in Tempe and his colleagues distinguished these basalts from other lunar rocks by measuring the tell-tale signature of TiO_2 in reflected ultraviolet light. Such measurements might also inform decisions about the location of future Moon bases, because TiO_2 can be broken down to supply oxygen.

The Hubble camera the team used has since broken, but the Lunar Reconnaissance Orbiter, set to launch in 2008, can try the same technique. Space telescopes are best for the job because Earth's atmosphere blocks much of the ultraviolet spectrum.

CELL BIOLOGY

A short fuse

Cell **130,** 165–178 (2007)
Researchers have uncovered part of the mechanism underlying autophagy, the process by which cells digest and recycle cytoplasmic proteins and organelles.

Yoshinori Ohsumi at the National Institute for Basic Biology in Okazaki, Japan, and his colleagues studied the role of a protein called Atg8 that joins with a lipid called phosphatidylethanolamine (PE). Formation of autophagosomes, the membrane-bound bodies that engulf material to be recycled, requires Atg8–PE.

Using artificial membranes, the researchers showed that an assemblage

of Atg8–PE molecules binds membranes together and allows them to partly fuse (a process known as hemifusion). Mutated forms of Atg8 that did not cause the membranes to cluster *in vitro* also failed to form autophagosomes *in vivo*.

NANOTECHNOLOGY

One step at a time

Small doi:10.1002/smll.200600721 (2007)

Silicon particles, the stickiness of which depends on pH, can be persuaded to assemble in a predetermined sequence, researchers in Japan have shown. They suggest that the technique might be used to build structures for microelectronics or biochips.

Hiroaki Onoe of the University of Tokyo and his colleagues made particles that had some attachment sites decorated with an organic film (A) and others with a smooth silicon surface (B). The B sites are sticky only at low pH, so particles mixed into a solution that is only weakly acidic pair up by binding at their A sites. When the pH is lowered, these pairs join up through their B sites.

To demonstrate the technique, the researchers made U-shaped particles that linked into X-shaped pairs; these pairs then assembled into a structure resembling an interlocked chain.

CELL BIOLOGY

Stuck back together

Science **37,** 242–245; 245–248 (2007)

A protein complex important for binding together pairs of chromosomes during cell division and DNA repair has been found to act independently of DNA replication. It was previously thought that the complex, called cohesin, could act only on replicating chromosomes.

Two research groups — one led by Camilla Sjögren of the Karolinska Institute in Stockholm, Sweden, and one led by Douglas Koshland of the Carnegie Institution in Baltimore, Maryland, and their colleagues — found that a wave of cohesin complexes formed around a break in a yeast chromosome, in addition to the binding of cohesin seen after replication. Furthermore, selectively damaging only one chromosome activated cohesin complexes on both damaged and undamaged chromosomes.

The results suggest that cohesin responds to DNA damage to help maintain chromosome integrity, although the mechanism for this is not fully understood.

ASTRONOMY

Star bright

Astrophys. J. **664,** L17–L21 (2007)

Space telescopes have looked back around 9 billion years to see a dense cluster of galaxies taking shape. Astronomers think that some of today's largest galaxies formed through the merger of smaller ones, and that this cluster may be a system on the brink of such a merger.

The cluster contains at least 12 massive galaxies. Patrick McCarthy of the Observatories of the Carnegie Institution of Washington in Pasadena, California and his colleagues estimate that the galaxies' total mass is nearly a trillion times that of the Sun. They suggest that several of the central galaxies may have merged within a few billion years of the light now observed leaving the system, perhaps forming what is known as a 'brightest cluster galaxy'.

PALAEONTOLOGY

Forerunner to roadrunner

Naturwissenschaften **94,** 657–665 (2007)

Palaeontologists have discovered 110-million-year-old footprints of a fleet-footed bird. The finding predates previously described avian runners by at least 50 million years.

The fossilized footprints, found in Shandong province in China, were originally reported two years ago but researchers initially thought the bird tracks resembled those of a modern shorebird. Martin Lockley of the University of Colorado, Denver, and his colleagues have since reanalysed the *Shandongornipes* tracks and found that the bird had feet more like a roadrunner's (pictured), with two toes pointing forwards and two pointing backwards.

T. BEAN/GETTY

Based on the spacing of the tracks and the projected height of the bird, the researchers estimate its speed to have been 8 kilometres per hour.

JOURNAL CLUB

Colin Prentice
QUEST, University of Bristol, UK

A theoretical biologist suggests that evolution makes plants more predictable.

The debate over how forests respond to rising levels of carbon dioxide has brought home to me how much spin even a dry journal article can contain.

In the mid-1990s, when the forest Free Air Carbon dioxide Enrichment (FACE) experiments began, I thought that we were poised to learn how trees really respond to CO_2. In these experiments, CO_2 is pumped over forests to simulate future conditions.

Unfortunately, years of data collection and scores of papers later, we still haven't reached agreement. Using the same data, researchers conclude that CO_2 either fertilizes forests or it doesn't (or the effect is small, or it goes away, or will soon go away...)

The situation would be helped if we had better theories of how trees might be expected to react to changes in their resources. It was refreshing, therefore, to encounter an elegant analysis of plant behaviour (O. Franklin *New Phytol.* doi:10.1111/j.1469-8137.2007.02063.x; 2007).

Plants, subject to selective pressure, have to optimize what they can. This is a basic principle of evolutionary biology, too often disregarded in experimental contexts.

Theoreticians have long known that an individual leaf in high CO_2 will maximize the amount of carbon it fixes — a measure of its growth success — if it lowers its nitrogen content to optimize the balance between photosynthesis and respiration.

Franklin extends this nitrogen optimization principle to the whole plant, a significantly more complex problem. His model predicts 83% of the variation in plant growth enhancement seen across FACE studies, explains the observed relationship between plant growth and canopy nitrogen content, and does much else besides. It is a welcome step forwards.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

SPECIAL REPORT

High noon in Libya

This week sees yet another crisis point in the Libyan case of six foreign health professionals sentenced to death on charges of injecting hundreds of children with HIV. **Declan Butler** traces the efforts of scientists to help establish the truth.

Rich Roberts didn't realize what he was getting into last October, when he decided to mobilize his fellow Nobel laureates to draw attention to a death-penalty case in Libya. Six medical workers — five nurses from Bulgaria and a Palestinian doctor — were charged with deliberately infecting more than 400 children with the virus that causes AIDS. Roberts, like many scientists, was shocked at how scientific evidence exonerating the medical workers had been ignored, and decided to do something about it.

For Roberts, a 1993 Nobel laureate in medicine or physiology, it was the start of a relentless commitment. Over the past nine months, he has had a string of meetings with top-level diplomats, and on 10 June he even flew to Libya for a late-night meeting at the Corinthian Hotel in Tripoli with Seif al-Islam Gaddafi, son of the Libyan leader Muammar al-Gaddafi, to try to help find a solution.

Roberts has also spent endless hours gathering an eventual 120 signatures from Nobel laureates — a record — for an open letter to Muammar al-Gaddafi. "A miscarriage of justice will take place without proper consideration of scientific evidence," warned the letter (see *Nature* **444,** 146; 2006). And it called on "the appropriate authorities to take the necessary steps to permit such evidence to be used in this case".

On 31 October, just days before it was published, Roberts hand-delivered the letter to Libya's ambassador to the United Nations, Attia Omar Mubarak. Their meeting in New York lasted an hour and a half. Mubarak was dismayed about the letter, arguing that the Islamic way was to try to negotiate a settlement between the accused and the injured. But, says Roberts, Mubarak did admit the possibility that the whole thing was an accident that had been seized on by prosecutors.

As *Nature* went to press this week, Libya's Supreme Council of the Judicial Authority was expected to annul or commute the death sentences of the six medical workers, a verdict that had been upheld by the country's supreme court just days before. The saga's anticipated ending is the result of months of careful negotiating between diplomats, informed and helped along by the input and advocacy of leading scientists.

Consideration of solid evidence is something the medical workers need badly. They were arrested in 1999, after an outbreak of HIV in more than 400 children at the Al-Fateh hospital in Benghazi; more than 50 of them have since died. The medical workers were initially charged with deliberately injecting the children as part of a plot by the US Central Intelligence Agency. Those charges were then dropped, with prosecutors now claiming that the medical workers used the children as guinea pigs to test a therapy in an illicit clinical trial.

In the midst of such spy-novel overtones, scientists have worked to inject credible evidence into the case. For instance, Vittorio Colizzi, an AIDS researcher at Tor Vergata University in Rome, Italy, testified at the medical workers' first trial along with Luc Montagnier, whose group at the Pasteur Institute in Paris discovered HIV. The scientists presented evidence that the infections were accidental, the result of a lack of safety precautions at the hospital. Other researchers, including Luc Perrin of the Geneva University Hospital in Switzerland, had reached the same conclusions independently. But the court threw their arguments out, on the basis that an investigation by Libyan doctors had reached the opposite conclusion.

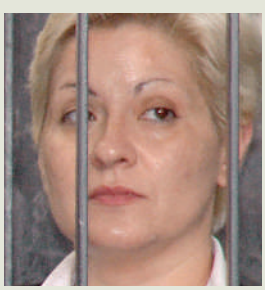
THE TRIPOLI SIX

M. TURKIA/AFP/GETTY

Ashraf Ahmad Jum'a

Valia Cherveniashka

Snezhana Dimitrova



Nasya Nenova

Kristina Valcheva

Valentina Siropoulo

P. PETROV/AP

Activist groups have shown their support for the medical workers sentenced to death.

L. LARBI/REUTERS

AIDS MEDICS IN LIBYA
Find all our coverage of this story online.
www.nature.com/nature/focus/aidsmedicslibya

For Colizzi, the stakes go beyond the death penalties. He fears that blaming the medical workers is part of wider denial of AIDS by Libya, which has a large population of migrant workers from sub-Saharan Africa who have HIV. Libya and many countries also need to face up to the problem of hospital-borne diseases and to introduce safe healthcare programmes, he adds.

United front

Last autumn, after an editorial in *Nature* called for activism (see *Nature* **443,** 605–606; 2006), scientific and human-rights groups including the American Association for the Advancement of Science, the Federation of the European Academies of Medicine, and the New York Academy of Sciences renewed public appeals and letter-writing campaigns to politicians. Robert Gallo, an HIV expert at the University of Maryland in Baltimore, and 43 other leading international researchers followed with a letter in *Science*.

Although most groups kept their actions to appeals, others were more hands-on. The Massachusetts-based Physicians for Human Rights campaigned publicly, and used its well-established political networks to quietly press the case. The International Human Rights Network of Academies and Scholarly Societies also worked behind the scenes.

Meanwhile, a team of European experts in the molecular phylogenetics of viruses decided to apply their expertise to the HIV sequences of the infected children being treated in Europe. The initiative of the group, who until then knew little about the case apart from what they had read in the news, was to provide crucial new evidence.

The initial phylogenetic analyses of the sequences confirmed epidemiological evidence that the infection had started at the hospital before the medical workers had started working there. As the retrial drew to a close, the researchers worked night and day to finish their analyses, suspending all other uses of their 40-processor cluster supercomputer to dedicate it solely to analysis of the Libyan sequences.

"Scientists have worked to inject credible evidence into the case."

Given the stakes, the team tested and re-tested their findings using multiple models. "We decided to throw the book at the data," recalls Oliver Pybus, an evolutionary biologist at the University of Oxford, UK. The results of every model were concordant; the start of the outbreak predated the March 1998 arrival of the medical workers. The paper was published online in *Nature* on 6 December, just before the scheduled court verdict (see *Nature* **444,** 836–837; 2006).

On 19 December, the court again handed down the death sentence to the medical workers. But if the findings had no effect on the court decision, they nonetheless had a major indirect effect by highlighting that scientific evidence had been ignored. The paper, coming on top of the Nobel letter and the rest of the scientific advocacy, catalysed an explosion of international outrage to the verdict, putting intense pressure on both Libya and the international community to renew efforts to find a way out of the crisis.

Then, on 1 January 2007, Bulgaria joined the European Union — a further turning point in the case, as it could now count on the diplomatic clout of the 27-nation body. In the months that followed, the most prominent role in talks was by diplomats from the European Commission and Britain, and from Seif al-Islam Gaddafi — a key intermediary through his charity, the Gaddafi Development Foundation. By contrast, US officials, although calling for Libya to exercise clemency, did not figure prominently behind the scenes.

Political influence

Events accelerated in early June, with visits to Libya by Benita Ferrero-Waldner, the European commissioner for external affairs; Tony Blair, the British prime minister at the time; and Frank-Walter Steinmeier, the German foreign minister. A possible deal that emerged was that the Supreme Court would uphold the death-penalty verdict, but that this would be quickly cancelled by a higher political body. The strength of the international medical and scientific advocacy gave diplomats additional grounds to push towards reaching a speedy conclusion.

The broad contours of the proposed deal remain as they have long seemed: humanitarian aid for long-term treatment of the infected children, which scientists emphasize is essential as the children too are victims of the tragedy. Families of the children, who have been told for years that the medics were guilty, are expected to receive US$1 million per child in compensation, funnelled through the Gaddafi foundation. Media reports have portrayed this as debt relief for money Libya owes dating back many decades. But the finer details of the diplomatic settlement remain shrouded in mystery, and it is far from clear as to where the money will come from.

For Roberts, the experience of this case has convinced him that scientists can do more in human-rights causes. Writing letters is useful, he says, but scientists can make a bigger difference if they engage personally with the diplomats and others involved. "We scientists can be much more effective if we are prepared to spend the time fighting for the issues in which we believe strongly." ■

Declan Butler

Deep science strikes gold after latest site is named

In the world of underground science, space is tight, and getting tighter. So scientists across the globe are welcoming a proposal for a new US facility that could help relieve the growing subterranean real-estate crisis.

On 10 July, the US National Science Foundation (NSF) announced that it had selected the abandoned Homestake gold mine near Lead, South Dakota, as the preferred site for a US$500-million Deep Underground Science and Engineering Laboratory. If fully funded, the mine will be developed into a sprawling underground campus — the deepest yet — where geologists, microbiologists and physicists can ply their trade.

"There is no more empty space."

It is physicists in particular who want the space, and who have been driving the push for the new lab. For decades, they have travelled to road tunnels and abandoned mines to build experiments that must be shielded from cosmic radiation. Only a handful of locations can host the searches, and many are becoming overcrowded, says Eugenio Coccia, director of the world's largest underground facility, the 180,000-cubic-metre Gran Sasso National Laboratory near L'Aquila, Italy. "There is no more empty space," he says.

In the United States, the situation is even worse, says Bernard Sadoulet, a physicist at the University of California, Berkeley. America's only major underground facility is at the Soudan mine in Minnesota, and Sadoulet, who co-chaired a review of underground science for the NSF, says his committee received around 80 letters of interest in the new lab.

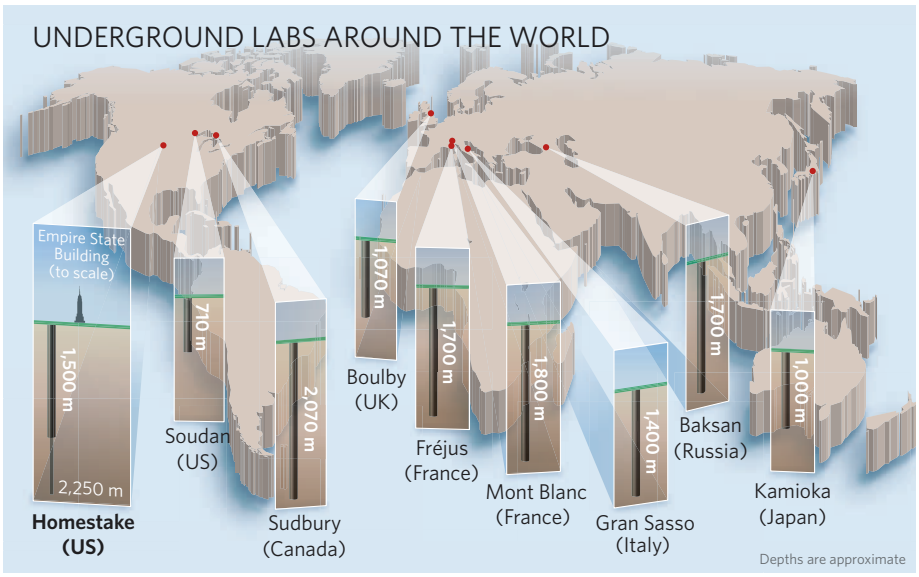
Then there is the problem of depth: the Soudan mine is 710 metres deep, but most of the newer experiments need to go deeper. Sadoulet's own experiment at Soudan probes dark matter — particles that interact only rarely but make up roughly a quarter of the mass of the Universe (see page 240). He says the next generation will be ready before Homestake and will be sited at the 2,070-metre-deep Sudbury Underground Laboratory in Ontario, Canada.

Such physics is being driven ever deeper underground. As particle accelerators increase in size and cost, rare-event physics of the sort that can be done only beneath thousands of metres of rock is seen as an increasingly attractive means of probing big questions. For instance, a phenomenon known as neutrinoless double-beta decay, an extremely rare event that occurs during the decays of some nuclei, could prove that neutrinos are their own antiparticle. Such a finding would have implications for the standard model of particle physics, and could explain why there is more matter than antimatter in the Universe today.

The search for these rare processes typically requires that detectors sit for years gathering just a handful of events. With each generation, the detectors get bigger and the experiments get longer. Thus, many existing underground laboratories are planning expansions. Proposals are on the table to expand the Fréjus laboratory in France in 2012, says Gilles Gerbier, a physicist at the French Atomic Energy Commission in Saclay. And the Kamioka Observatory in Japan, which hosts the massive Super-Kamiokande neutrino detector, is also digging new spaces for a dark-matter search and double-beta-decay experiment, according to Yoichiro Suzuki, the observatory's director.

The need for increased sensitivity of such experiments also sends them deeper, because each successive metre of rock gives more effective shielding from disruptive cosmic rays. The world's two deepest laboratories, Mont Blanc in France and the Sudbury mine in Canada (see graphic), are cramped. Even after an expansion at Sudbury planned for completion by 2008, there will be room for only four large experiments, says Arthur McDonald, a physicist at Queen's University in Kingston, Ontario. "We've had more letters of interest than we've space to house the experiments," he says.

Homestake would also provide opportunities for geologists and microbiologists. Geologists could use it to study how rock behaves under pressure; such information may help to understand earthquakes. Meanwhile, microbiologists see the lab as an opportunity to study organisms that live far beneath Earth's surface. "Most of these organisms don't depend on oxygen to survive," says Tullis Onstott, a geomicrobiologist at Princeton University in New Jersey. What nutrients they need, and how they obtain them, could provide clues to how life began, he adds.



MÖBIUS STRIP UNRAVELLED
Mathematicians solve 75-year-old mystery of infinite loop's shape.
www.nature.com/news

M. C. ESCHER

D. LAMMERS/AP

The abandoned Homestake gold mine will host the world's deepest underground lab.

If approved, the Homestake lab will have campuses at 1,500 and 2,250 metres below the surface, with cavities 50–60 metres in diameter. That would be big enough to handle detectors for the most ambitious searches, says Kevin Lesko, head of the Homestake collaboration and a physicist at Lawrence Berkeley National Laboratory in California. "I'm very excited," he adds.

The selection of Homestake caps a long and highly politicized process. It was first put forward as a candidate in 2001, quickly winning the backing of powerful local politicians such as Senator Tom Daschle, who was then Democratic minority leader. In 2005, the NSF announced that Homestake and a Colorado mine were the finalists for hosting the underground lab, but protests from losing teams caused the process to be reopened.

Even now, there is no guarantee that the Homestake lab will be built. Local billionaire T. Denny Sanford, together with the state of South Dakota, have pledged some $100 million for an interim site at 1,500 metres, but the deeper facility will require NSF construction money. At present, the agency has approved just $15 million for a three-year, detailed design study. To win full funding, the design must go before the independent National Science Board, where it will compete with other large projects.

Even so, researchers are pleased that the first steps have been taken, and are hopeful that the lab will be built. "It's clear where the science is going," says Sadoulet. "The frontier is deep." ■

Geoff Brumfiel

Russia pins its hopes on 'nano'

MOSCOW

In what could be the biggest windfall for science since the collapse of the Soviet Union, the Russian parliament last week gave the green light to a massive US$7-billion investment in nanotechnology over five years. The Russian government hopes the programme will make the country a world leader in nanoscale technologies with a wide range of military and civilian uses.

However, the move has been criticized as poorly prepared and unlikely to yield results.

Nano-devices, designed from single atoms and molecules, are predicted to have applications in fields as diverse as consumer electronics and biomedicine. All research and development activities will be coordinated by Rosnanotekh, a new tax-exempt body with far-reaching freedom to set up institutes, put work out to tender and commercialize results.

But no details have been announced about the precise structure, goals and content of the initiative. It is unclear, for example, how projects will be selected for funding.

Some Russian scientists, sceptical about fair allocation of funds, have given the announcement a lukewarm response. The country has hardly any competence in nanotechnology, they say. And given the widespread absence of efficient quality control in Russian science funding, many fear the scheme will be poisoned by corruption.

"Our government just doesn't understand anything about science," says one high-level Russian physicist who asked not to be named. "They think if they throw enough money at it they'll get some nice exploitable results in return. But we don't even have the experts."

"Lack of transparency and programme abuse are the usual Russian dangers."

The programme is the brainchild of Russian President Vladimir Putin, who is keen to reduce the country's dependence on oil and gas.

Putin recently compared the importance of nanotechnology to that of nuclear science. He is said to have secretly recruited Mikhail Kovalchuk, the director the Kurchatov Institute in Moscow, to head Rosnanotekh.

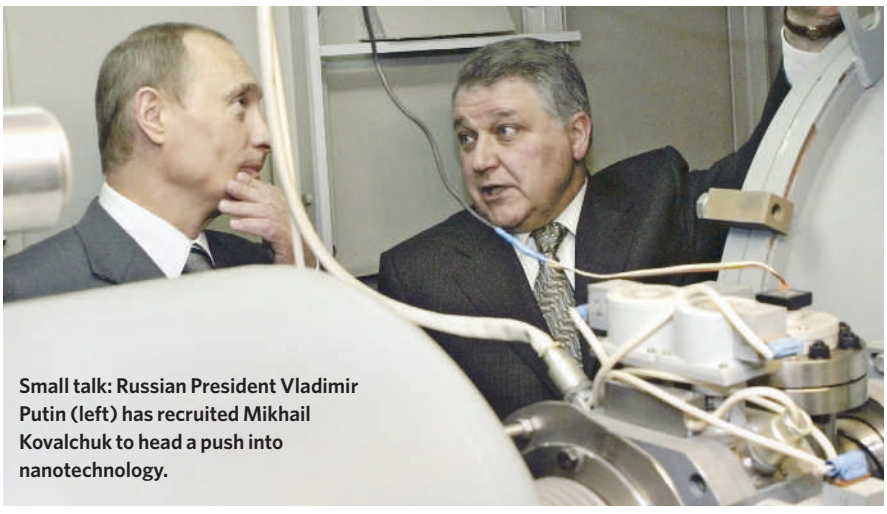
Kovalchuk, who is not an expert in nanotechnology, is the brother of Yuri Kovalchuk, a banker and businessman with close ties to Putin.

The independent Russian media has poured scorn on Russia's foray into what some call the "banano" technology business. "Lack of transparency and programme abuse for personal goals are the usual Russian dangers," says former science minister Boris Saltykov, an expert in science management.

"Risks do exist," agrees Alexander Nekipelov, vice-president of the Russian Academy of Sciences in Moscow. "But the money involved is so huge that scrutiny will be very good this time."

The academy is keen not to be bypassed by the programme, for which the government has set aside more funds than the entire academy receives. In a move that critics say violates its own rules, in June the academy leadership appointed Kovalchuk, who is not a full member, as acting vice-president for nanotechnology. ■

Quirin Schiermeier



M. KLIMENTYEV/RIA NOVOSTI/PRESIDENTIAL PRESS SERVICE/AP

Small talk: Russian President Vladimir Putin (left) has recruited Mikhail Kovalchuk to head a push into nanotechnology.

ON THE RECORD

"We can categorically state that we have not released man-eating badgers into the area."

UK military spokesman Major Mike Shearer denies rumours that British forces had sowed panic in Basra, Iraq, by unleashing ferocious honey badgers (pictured). The badgers, slightly larger and scarier than the British woodland version, are in fact indigenous to the Middle East.

AFRIPICS.COM/ALAMY

SHOWBIZ NEWS

We will doc you

Former Queen guitarist Brian May has finally completed the astrophysics PhD he abandoned 33 years ago after becoming distracted by international superstardom. He has submitted his thesis to Imperial College, London, and is set to become Dr May in May next year. Rock 'n' roll.

ZOO NEWS

Bear burnout

Berlin Zoo's Thomas Dörflein, who found fame as the keeper of Knut, the world's cutest bear™, has bowed out of the limelight. He has retired from public play sessions with the young polar bear, who now weighs 50 kilograms and is becoming a bit of a handful.

ROBOT NEWS

Jesus-bot

Engineers at Carnegie Mellon University in Pittsburgh, Pennsylvania, say they have created the first robot that can walk on water. Unfortunately, it can carry a load of only 9.3 grams.

ZOO/ROBOT NEWS

Lamprey power

Roboteers led by Ferdinando Mussa-Ivaldi of Northwestern University in Evanston, Illinois, have created a cyborg that uses a lamprey brain to control a light-seeking robotic disc. They hope it could lead to prosthetic aids for people paralysed by stroke or motor neurone disease.

Sources: BBC, The Times, CNN, PhysOrg, Small Times

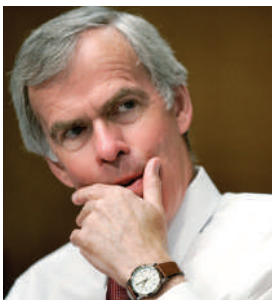
SIDELINES

US proposal for carbon cuts offers compromise

Legislation to cut carbon emissions has traditionally received little support in corporate boardrooms and union halls, but this may soon change. Several large utility companies are among those backing a new proposal offered in the US Congress on 11 July by Senators Jeff Bingaman (Democrat, New Mexico) and Arlen Specter (Republican, Pennsylvania).

The Bingaman–Specter proposal is the latest of several major climate bills now under consideration by Congress (see table). Some observers see this one as setting the tone for a compromise package aiming to bring together competing interests to fight climate change. The 'Low Carbon Economy Act' would require the United States to reduce its carbon output to 2006 levels by 2020 and to 1990 levels by 2030. (The Kyoto Protocol on climate change, which the United States has not ratified, calls for cuts below 1990 levels by 2012.) Further reductions, to at least 60% below 2006 levels by 2050, are contingent upon cuts being made by other countries.

Companies such as Duke Energy, one of the nation's largest utility suppliers, back the new bill because it includes a provision allowing carbon emitters to buy extra allowances at a set cost. This would provide a 'safety valve', ensuring a stable price for emissions, and keeping the companies' future costs at a foreseeable level.



A. WONG/GETTY

Jeff Bingaman's climate bill has industry backing.

The bill also has the support of the union group AFL-CIO, which has traditionally challenged climate-control legislation on the grounds that it would drive jobs overseas, and key Republican legislators — such as Lisa Murkowski and Ted Stevens, both senators from Alaska — who have been sceptical of other climate proposals. "The other bills are more aggressive and less realistic," says Frank Maisano, a spokesman in Washington DC for Bracewell & Giuliani, a law firm representing many of the fossil-fuel industries. "They're all show and no go."

But environmental advocates say that pricing extra allowances at a set cost will weaken the nascent American carbon market. "I think it's unlikely to achieve absolute reduction if the safety valve undercuts the programme," says Vicki Arroyo, director of policy analysis at the Pew Center on Global Climate Change in Arlington, Virginia. "It's more like an escape hatch."

Get practical, urge climatologists

British experts have criticized the focus of current climate projections. They say that scientists should shift from models that predict what will happen many decades from now, and concentrate instead on shorter-term forecasts that will aid policy-makers, businesses and the public.

Climate models such as those used in the Intergovernmental Panel on Climate Change (IPCC) reports have been instrumental in convincing the world that climate change threatens ecosystems and human societies, but they do not provide much practical guidance. "We may not be providing what we possibly could," says Peter Cox, a climate modeller at the University of Exeter, UK, and former chair of climate-system dynamics at the UK Met Office.

Cox and his colleague David Stephenson, also at the University of Exeter, published their argument last week (P. Cox and D. Stephenson *Science* 317, 207–208; 2007). "The IPCC has nailed many old questions," says Cox. "It's a done deal, so we had better move on."

A key question is how to make climate-change models socially relevant. Cox and Stephenson propose having climate forecasters shift their attention to around 2050, rather than trying to predict farther into the future. This would effectively mean that the timescale of climate predictions would match that over which long-term policy and business planning is carried out.

The authors note that climate models are least uncertain for between 30 and 50 years from now. Shorter-term predictions will be less accurate because of

CLIMATE CHANGE
Find all our stories on climate in one place online.
www.nature.com/news/infocus/climatechange.html

CLIMATE BILLS IN CONGRESS

	Sponsors				
Provisions	Bingaman–Specter	Lieberman–McCain	Sanders–Boxer	Feinstein*	Kerry–Snowe
Emission target by 2050	60% below 2006 levels — provided other countries play ball	60% below 1990 levels	80% below 1990 levels	Cut expected levels for 2020 by 25%; 1.5% annual reductions thereafter	65% of 2000 levels
Carbon allowance	53% to industry; 24% for auction; 9% to states; 14% to others	Allowances distributed across sectors and to a new 'Climate Change Credit Corporation'	Awarded to those most affected by transition to a carbon-free economy	Allowances based on means of electricity generation	To be determined by the president
Technology support	Creates fund for research into low-carbon technologies and vehicles. Supports carbon capture and storage	Climate Technology Finance Board backs public–private research partnerships. Climate Change Credit Corporation supports low-carbon technologies	Grants for carbon capture and storage projects.Recommends boosting R&D for low-carbon technologies by 100% a year for a decade	Climate Action Trust Fund established to commercialize new low-carbon technologies	Recommends boosting R&D by 100% a year for a decade. Creates programme to assist with adaptation to climate variation

*Applies to electricity sector only
Source: US Senate

Under the bill, 53% of carbon allowances would be handed out to utilities, manufacturers and other carbon-producing industries. From 2017 on, an increasing proportion of allowances would be auctioned off, generating billions of dollars for green technology and climate-change adaptation. The safety valve would start at $12 per tonne of carbon dioxide emitted, and would increase each year at 5% above inflation.

The Bingaman–Specter bill also packs in several features that don't appear in other proposed legislation, says Jonathan Pershing, a climate expert at the World Resources Institute, an environmental think-tank in Washington DC. It provides additional allowances for companies that invest in carbon capture and storage, and spells out how the government will divvy out carbon allowances, including a share for states to distribute. "None of the other bills has this," Pershing says.

The major sticking point for environmental groups is the bill's safety valve. If the market price for carbon exceeds the safety valve, emitters can instead buy allowances — essentially a carbon tax. The cost of credits is likely to surpass the limit early on and disrupt the carbon market, says Arroyo.

Such a feature would also complicate US participation in an international carbon-trading market. European governments are unlikely to allow companies to purchase American offsets directly, but a less formal link-up based on options trading could emerge, says Pershing. The European carbon market, which opened in 2005, has got off to a shaky start. Prices for a tonne of carbon plummeted from €31 (US$43) to €12 in April 2006, when leaked emissions data revealed that several nations hadn't used up their allotted credits.

Jeff Holmstead, former head of the Environmental Protection Agency's Office of Air and Radiation and now also with Bracewell & Giuliani, says a safety valve is needed because technologies to make deep cuts in carbon emissions are not yet available.

Larger questions about effects on the domestic economy and the likely migration of carbon-producing industries abroad mean climate legislation has no prospect of passing soon, Holmstead argues. He is confident that there will be insufficient votes supporting it in either the House of Representatives or the Senate "until there's a much better understanding of what it will mean".

Those pushing for a strong climate bill are more optimistic. Congress is likely to act soon, says Pershing, given the growing pressure from the US public to address global warming. "The question is not whether, but when," he says.

A subcommittee led by Senators Joe Lieberman (Independent, Connecticut) and John Warner (Republican, Virginia) is expected to work out a compromise climate bill that is likely to reach the Senate floor in coming months. ■

Ewen Callaway

uncertainty over initial conditions. Changes that will transpire over the next three decades are essentially "already in the system", says Cox. For predictions more than 50 years in the future, uncertainty levels in the models increase because no one can accurately forecast the level of carbon dioxide emissions resulting from human activity.

J. WOOD/GETTY

Out of water: practical solutions are needed to the problems of climate change.

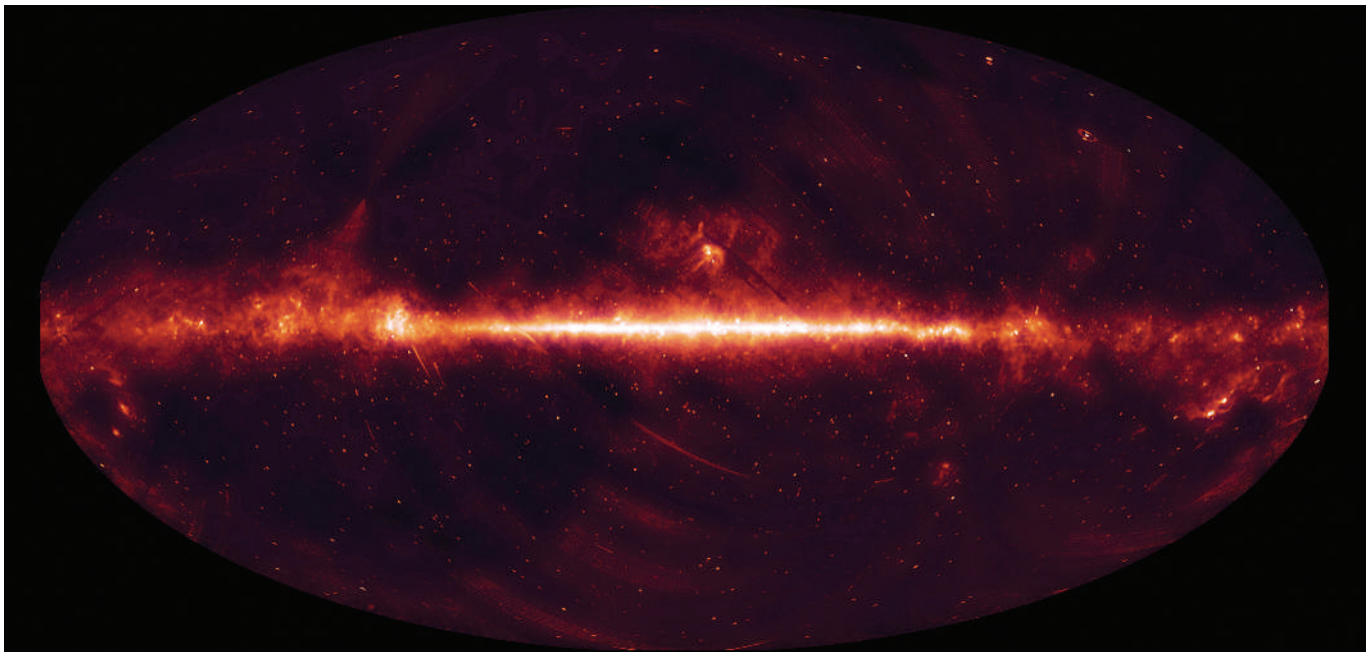
Between 30 and 50 years away is thus a sort of sweet spot in which to target policy planning, Cox says. Mitigation policies and plans for associated socioeconomic factors, such as economic growth, energy use and technology needs, could be developed with that time frame in mind.

Some of those involved in the IPCC process do not disagree in principle, but say the inherent uncertainty of climate models will always make forecasting difficult. "Focusing on what 'should be' is a worthy goal going forward, but not a panacea for the uncertainty problem," says Cynthia Rosenzweig, a climate modeller at NASA's Goddard Institute for Space Studies in New York, and a coordinating lead author of the most recent IPCC review, published this year.

Another report released last week also criticizes the "big gap" between how climate scenarios are currently used and their "potential contributions" to policy making. The report was co-authored by Rosenzweig, and is the second in a series by the US Climate Change Science Program.

The current generation of climate scenarios is still useful for resource managers to guide their preparation for climate change over the next few decades. Planning for such change is becoming "mainstream" in water-management systems, coasts and human health care, says Rosenzweig. ■

Quirin Schiermeier



JAXA/ISAS/LIRA

SNAPSHOT
Seeing red

The entire sky's infrared emissions are captured in this image released last week by the Japan Aerospace Exploration Agency (JAXA) and the European Space Agency (ESA). The plane of the Milky Way galaxy appears as a bright strip running through the middle, with the black hole at its heart emitting strongly. The bright spot at the lower right is the neighbouring galaxy the Large Magellanic Cloud.

This map was assembled from thousands of images taken by JAXA's AKARI satellite, which has been in orbit since February 2006. "What you are seeing now is the result of the comprehensive coverage of the entire sky, the result of a year's scanning," says team member Chris Pearson of ESA, who is based in Japan.

The latest map has a higher resolution and sensitivity than the whole-sky infrared map compiled by the Infrared Astronomical Satellite (IRAS) in 1984. Stars that were previously blurs are now distinct entities.

AKARI has also surveyed colder galaxies than IRAS was able to detect. It will keep collecting data until it runs out of the liquid helium needed to keep its detectors cold — which is expected to happen in early September 2007. ■

David Cyranoski

Science watchdog baulks at merger

Members of an influential parliamentary committee, which serves as a watchdog on UK science policy, are protesting over a planned government reorganization that they fear would subsume it into a larger committee.

Under the plans, the Select Committee on Science and Technology would become part of the newly created Department for Innovation, Universities and Skills, which is responsible for science now that Gordon Brown is prime minister (see *Nature* **448,** 7; 2007). Members worry that the new committee would be mainly concerned with university administration.

The science select committee, consisting of members of the three main political parties, is well known for overseeing the government's policy on issues ranging from illegal drugs to the proposed national identity-card scheme. Earlier this year, it also successfully campaigned to overturn the government's proposed ban on hybrid embryo research.

Given the current high public profile of science, now is "the wrong time to downgrade or reduce the scrutiny of cross-cutting science issues within parliament", committee chair Phil Willis wrote in a letter to Geoff Hoon, the government's chief whip. "The strong view amongst the science community is that such scrutiny is best carried out by a select committee with a clear identity and a clear mission."

"It's a pity that the name 'science' has been lost because it will reduce its profile. That's the danger."

Another committee member, Evan Harris, added: "The problem is that a committee that covers universities and skills is not going to have time to scrutinize science across the government." It would be preferable, he says, for a dedicated, independent science committee to continue with the remit of the current one.

Harris raised the possibility in parliamentary questions on 12 July. House of Commons leader Harriet Harman replied that "discussions are ongoing". Harris told *Nature* that the committee's ideal position would be to remain as an independent entity, rather than become a subcommittee within the new department.

Critics of the government's reshuffle have pointed out that the new department, although responsible for overseeing science, does not even feature the word 'science' in its title. "It's a pity that the name 'science' has been lost because it will reduce its profile. That's the danger," says Harris.

Previously, scientific research came under the remit of the Office of Science and Innovation, within the Department of Trade and Industry. ■

Michael Hopkin

See Editorial, page 226.

Threatening letters rattle evolutionary biologists

Evolutionary biologists at the University of Colorado in Boulder are on edge after receiving a number of threatening letters.

"I charge you and your devilutionist colleagues with being the source of every imaginable evil," read part of an e-mail sent to biologist Michael Grant, and posted on the science blog The Panda's Thumb. Grant says 8–10 colleagues have received similar e-mails, letters and packages in recent days.

The letters are reportedly signed "Michael Korn". A website that seems to be written by the same alleged author describes him as a born-again Christian in Denver, Colorado. The Michael Korn who runs this website denies any knowledge of the letters or e-mails.

"In general, people are worried," Grant says. "We have one faculty member and one graduate student who are scared to go into the department." The university's Police Department is investigating.

Budget promises boost for German science

Germany's coalition government has proposed giving the Ministry of Education and Research a near-8% rise in its budget for 2008. If approved, the ministry would receive €9.2 billion (US$12.7 billion) next year, a rise of €670 million on this year.

Energy research and environmental science, particularly climate change, would get €336 million of the extra funds, says research minister Annette Schavan. The Max Planck Society, which runs 80 basic-research institutes, and the DFG, Germany's main grant-giving agency, will each see a 3% rise in their budgets.

The increase would raise Germany's science spending to 2.7% of its gross domestic product, close to the level of 3% envisaged by the European Commission for the European Union by 2010. The budget will be debated this autumn by the German parliament, which will vote on it in November.

US draws up short list of sites for bioweapons lab

Five sites are now competing to host a $450-million biosecurity complex, the US government announced last week. Selected from a group of 18, the finalists for the National Bio and Agro-Defense Facility (NBAF) include sites in Texas, North Carolina, Mississippi, Georgia and Kansas.

The Department of Homeland Security expects to make a final decision in late 2008, and the NBAF will open in 2013. The facility will do research on potential bioweapons and replace the 50-year-old Plum Island Animal Disease Center in New York.

Not making the cut was Texas A&M University in College Station. Earlier this month, the US government halted biodefence research there after reports that workers had been exposed to two potential bioweapons. The university, which had been one of the 18 sites bidding to host the new lab, reported the incidents more than a year after they occurred (see *Nature* 448, 13; 2007).

Australia joins Europe's biology laboratory

The European Molecular Biology Laboratory (EMBL) has offered Australia associate membership. The seven-year arrangement begins in January 2008 and aims to encourage scientific exchange between Australia's top research centres and EMBL's five European laboratories.

M. B. HANSEN/EMBL-PHOTOLAB

The European Molecular Biology Laboratory is strengthening its links with Australia.

EMBL, a non-profit research organization with 19 member countries, pointed to Australia's strengths in stem-cell research and medical epidemiology as reasons for the offer. The Australian government and several universities will together pay the membership fee. An EMBL representative would only say that the financial contribution is "significant". In return, Australia will be able to send scientists to train at EMBL and will have access to the labs' research equipment.

Only European countries, and Israel, can become full members of EMBL. Australia is the organization's first associate member. EMBL says that other countries have expressed interest in associate membership, and discussions are ongoing.

US House backs more stringent drug regulation

The US House of Representatives has overwhelmingly passed a bill that would give the Food and Drug Administration (FDA) unprecedented power to police the safety of drugs after they have been approved and gone on sale.

The 403–16 vote on 11 July follows the US Senate's near-unanimous approval of a similar measure in May (see *Nature* 447, 247; 2007). Both bills allow the agency to insist on, rather than negotiate, drug-label changes, and let the FDA order further clinical studies once a drug has been marketed.

But a disagreement is brewing and may emerge when House and Senate lawmakers meet to resolve differences between the two bills later this summer. Senator Edward Kennedy (Democrat, Massachusetts) has vowed to incorporate language making it possible for the FDA to approve copycat versions of biological drugs (see *Nature* 447, 629; 2007). The key House lawmaker in the bill, John Dingell (Democrat, Michigan), is opposed to this provision.

Because the bill is replacing a law that expires on 30 September and that provides more than $300 million in funds to the FDA, lawmakers are under pressure to pass it quickly to avoid major layoffs at the $1.9-billion agency.

The calf that came in from the cold

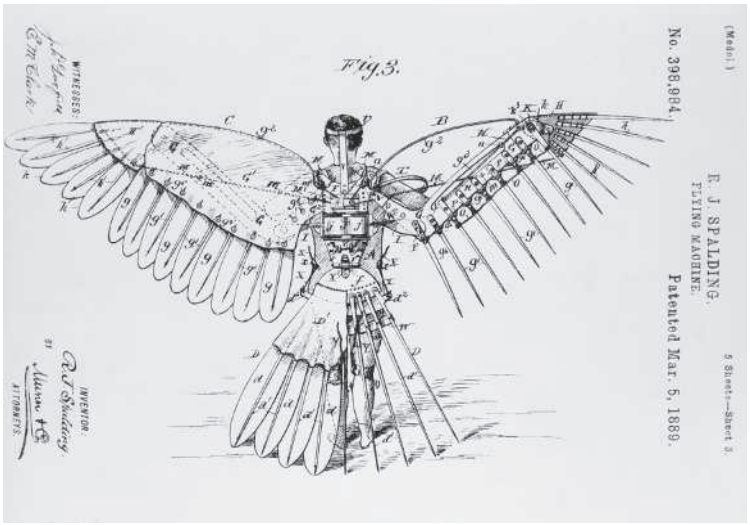
S. CHERKASHIN/AP

Meet Lyuba, a 4-month-old woolly mammoth found in the melting Siberian permafrost. Named after the wife of the reindeer breeder who discovered her, the mammoth emerged from at least 10,000 years of deep freeze, weighing 50 kilograms and retaining an unprecedented amount of her soft tissue. One of maybe five calves ever found, this specimen could have fetched millions of dollars on Russia's black market had it fallen into the wrong hands, says Larry Agenbroad, of the Mammoth Site in Hot Springs, South Dakota. Instead, Lyuba is headed for Jikei University School of Medicine in Tokyo, where researchers plan to perform computerized tomography scans on the calf.

BUSINESS

Patent examiners call in the jury

The US Patent and Trade Office has cracked open the door on its normally closed patent evaluation process. **Heidi Ledford** looks at how its peer-review project is faring.



BETTMANN/CORBIS

R. J. Spalding's 1889 patent for a flying machine: arguments over prior art haven't got any simpler since.

Microsoft patent application 11/296194 suggests a method for distributing music files that ensures copyright holders get their royalties, and started out the way most patents do. An application was filed to the US Patent and Trademark Office in December 2005, and sat there, unresolved, for more than a year. But last week, 11/296194 took a new path. The application was posted on a website, and visitors were asked to submit opinions and evidence to answer the question: is it new?

Within two days of posting the patent, there were two responses. "How could anyone consider this non-obvious?" wrote a computer technician from Liberty, Montana. A Hewlett-Packard software engineer disagreed: "I believe the application has merit."

Welcome to 'Peer to Patent', the patent office's pilot project launched on 15 June to bring patent evaluation to the masses (see www.peertopatent.org). The debate over the Microsoft patent is just beginning; registered visitors to the site will have 16 weeks to submit comments. The project aims to help the overstretched, backlogged and beleaguered US patent office search for 'prior art' — evidence that a patent's claims have already been patented or are already in common use. The hope is that tapping the expertise of hundreds of reviewers will improve the rate of prior-art discovery, speeding up the process for patent examiners while creating solid patents more likely to withstand future litigation.

Peer to Patent is a pet project of New York Law School's Beth Noveck, who sees it as a step towards bringing an open-source approach to otherwise closed government decisions. The project has now been in operation for five weeks and already has 1,000 registered users, eight patents under consideration and more than 30 submissions of evidence showing prior art. It has well-heeled sponsors, including General Electric, IBM, Microsoft and Hewlett-Packard, who hope it can help to unclog the process for granting information-technology-related patents. Peer to Patent has also piqued the interest of the UK Intellectual Property Office, which plans to launch a similar programme in the next year or so, Noveck says.

Beth Noveck: wants to open up patent decision-making.

John Doll, the US commissioner for patents, says he is optimistic about scaling up the programme, and sees expansion to include the biotechnology sector as a logical next step. "That would be a natural outgrowth," he says. "In the sciences you'll see a lot of interest."

Observers say that Peer to Patent has got off to an auspicious start. But although most welcome the programme as a sign that the US patent office is open to reform, some wonder how much of an impact it can really have. "It is an interesting experiment that seems to me worth trying — although I would be very surprised if it proved to be robust enough to affect the patent system more than marginally," says Dan Burk, a specialist in patent law at the University of Minnesota in Minneapolis.

The United States has traditionally opposed opening patent examinations to the public, citing the possibility of influence from competing interests or of overburdening the patent examiner with irrelevant claims, says Stephen Kunin, a former deputy commissioner at the US patent office, now at Oblon, Spivak, an intellectual-property firm based in Alexandria, Virginia. The Peer to Patent programme has built in several safeguards against petty interference in the process, however.

Suggestions of prior art must be accompanied by solid evidence, and the registered users are asked to agree to or reject a submission's addition to a top-ten list. Only the ten strongest suggestions of prior art are then forwarded to the examiner at the patent office. The applicant can view comments, make preliminary amendments and ask for an interview with the examiner before action is taken on the application.

For now, the programme is limited to 250 computer software and hardware patent applications. Some worry that it will be difficult to scale it up to handle a more significant caseload. "If it works for 100 patents, can we expand it to anything like a major subset of the 400,000 filed each year?" asks Mark Lemley, director of the law, science and technology programme at Stanford University, California.

Others note that biotechnology or pharmaceutical companies are far less enthusiastic about speeding up the patent process than are computer companies (see *Nature* **437,** 1230; 2005), and doubt if they will embrace Peer to Patent. Biotech companies, in particular, rely heavily on their patents in their business model, and have been resistant to radical changes to the patent system. "I would suspect they would be much more reluctant to participate," says Arti Rai, a law professor at Duke University in Durham, North Carolina.

It will take time to establish if the pilot project is speeding up approval and improving patent quality. Will the skills of hundreds of unpaid Peer to Patent volunteers match those of patent litigators once a patent is approved? "It's when somebody's willing to spend $2 million to $3 million on a lawsuit that people get really good at finding prior art," observes Burk. ■

D. PARKINS

Welcome to the dark side

Physicists say that 96% of the Universe is unseen, and appeal to the ideas of 'dark matter' and 'dark energy' to make up the difference. In the first of two articles, **Jenny Hogan** reports that attempts to identify the mysterious dark matter are on the verge of success. In the second, **Geoff Brumfiel** asks why dark energy, hailed as a breakthrough when discovered a decade ago, is proving more frustrating than ever to the scientists who study it.

We're underneath 1,400 metres of Italian mountain, walking through cavernous halls that lead from a 10-kilometre-long road tunnel. The scientists working within the Gran Sasso National Laboratory near L'Aquila seem ant-like in scale against the backdrop of vast metal spheres, towers and scaffolding that house their underground experiments. Physicist Elena Aprile is hurrying the group along, pointing out one project after another. She stops to take a photo of one, exclaiming at its size. We finally reach Aprile's XENON10 experiment, which is tucked away at the end of a small side tunnel. This is the project into which Aprile has poured her energy over the past few years, one of several experiments at Gran Sasso and around the world that are waiting for a passing piece of 'dark matter' to show itself.

Once upon a time, waiting for new particles to reveal themselves was a major endeavour. Scientists in the 1940s would also head to the mountains — to their tops, not to underground caverns — carrying emulsion-covered plates to capture strange new cosmic rays. But as particle accelerators became more powerful, physicists became adept at making their own novelties, and lying in wait for chance discoveries fell out of fashion. In this, dark-matter searches are something of a throwback.

They are a reminder of the past in another way, too. Ever more powerful accelerators require ever vaster detectors and ever larger teams of people to make sense of their output. The Large Hadron Collider (LHC) under construction at CERN, the European particle-physics laboratory just outside Geneva, will cost €3 billion (US$4.1 billion) and is the work of thousands of scientists and engineers. The XENON10 detector is run by just 30 scientists, and that's part of its attraction. "It's a last chance to do physics like it used to be done," says Aprile.

In the hunt for dark matter, a small team can make a big difference. XENON10 has steamed ahead of older collaborations to become the most sensitive detector for a category of dark matter called weakly interacting massive particles, or WIMPs. Other collaborations are keen to wrest the lead back, and over the next two to five years, sensitivity records look set to fall repeatedly. "A few years ago, I would have been surprised if dark-matter detectors had found a WIMP," says Leszek Roszkowski, a theorist from the University of Sheffield, UK, on sabbatical at CERN. "In a few years' time, if our ideas are correct, I will be surprised if they don't."

The dark titans

The first hints of dark matter came in the 1930s, when astronomer Fritz Zwicky spotted something odd about the behaviour of galaxies in the Coma cluster. His measurements of the galaxies' velocities suggested that the cluster was held together by more mass than he could see. He wrote: "If this is confirmed, we would arrive at the astonishing conclusion that dark

matter is present with a much greater density than luminous matter."

Cosmologists now believe that dark matter provides the scaffolding around which all other cosmic structures, from galaxies to galaxy clusters, superclusters and more, have taken shape. Astronomers are building big telescopes that can map its distribution in the heavens (see 'The search for structure', page 244). But this dark stuff cannot be the everyday matter of which stars, gas clouds and planets are made. Detailed measurements of the microwave radiation left over from the Big Bang suggest that such ordinary matter makes up just 4% of the Universe. The rest is thought to be divided between dark matter — outweighing normal matter by five to one — and a strange repulsive force dubbed dark energy (see 'A constant problem', page 245).

"If you look at the history of the Universe, it's been the battle of the two dark titans," says Michael Turner, a cosmologist at the University of Chicago, Illinois. "For the first 10 billion years, dark matter reigned, and it shaped all the structure in the Universe, and then, about five billion years ago, dark energy took over, shut off the formation of structure and got the Universe accelerating."

The idea that dark matter might not just be dark but fundamentally different from other matter gained ground in the 1970s. Planning for underground detectors similar to Aprile's began in the 1980s; but the field has heated up only recently as the sensitivity of WIMP detectors has improved — and as competing experiments have emerged to attack the problem from other angles.

Producing particles

Most of the particles that theorists have suggested as possible WIMPs are massive — at least 100 times the mass of a proton. Their size has kept them beyond the reach of particle accelerators, but the LHC could well change that, and produce in the lab what the dark-matter detectors have so far failed to capture in the field.

While researchers at the LHC have a new collider to tackle the problem, astronomers are taking yet another approach. Some of the WIMP particles that theorists are fond of might give off distinctive bursts of γ-rays or other odd signatures when they interact with each other; satellites and telescopes are now looking for such signals.

But the scientists trying to catch a piece of dark matter using their underground experiments still hope to get there first. "If you don't get your act together now," says Aprile, "the feeling is that you're going to be too late." Even those who have worked at CERN in the past, such as dark-matter researcher Bernard Sadoulet, now at the University of California, Berkeley, hope for a return on their investment in direct searches. "Of course after putting 20 years of my life into this thing, I would like to see it first," says Sadoulet.

Any of these experiments could individually provide evidence for a dark-matter particle. To date, there has been only one claim of direct particle detection, and that remains controversial (see 'Contested results'). So multiple lines of evidence will be essential for scientists to claim with confidence that they have discovered a new ingredient of the Universe — especially because that particle might point to a new framework of physical laws. "There's never been a more fun time to wonder about dark matter," says physicist Max Tegmark of the Massachusetts Institute of Technology in Cambridge. "It really feels like we are on the brink. There are these different roads to dark matter and they are all on the verge of coming through."

No one knows what dark matter is, but they know what it's not. It's not part of the 'standard model' of physics that weaves together everything that is known about ordinary matter and its interactions. The standard model has been hugely successful, but it also has some problems, and in trying to fix these, theorists have predicted hordes of new fundamental particles. At first, these hypothetical particles were viewed as unwelcome

J. HOGAN

Elena Aprile hopes her underground detector will catch dark matter.

CONTESTED RESULTS

Has dark matter been seen already? The DAMA collaboration based in Italy claims that an annual oscillation in the number of events registered by its underground detector is a dark-matter signal, but others are sceptical. "It is getting more and more difficult to reconcile DAMA with all the other results in the field," says Bernard Sadoulet, spokesperson for the CDMS, a competing dark-matter detector in the United States.

Over the course of a year, the number of dark-matter particles hitting Earth is expected to vary because the planet's velocity with respect to the rest frame of the Galaxy (and that of the Galaxy's dark matter) changes. In the summer, Earth moves in the same direction as the Sun does about the Milky Way's centre, so their velocities add together; in the winter these motions are opposed.

The DAMA experiment, based at the Gran Sasso National Laboratory near L'Aquila, Italy, looked for this effect between 1995 and 2002. Sure enough, the detector registered more particles hitting its 100-kilogram sodium iodide target in the summer than in the winter, and the team concluded that DAMA was seeing dark matter.

The problem is that other detectors searching for dark matter have failed to see any particles with the properties they would expect from the DAMA results. Critics of the experiment worry that other factors that could give a seasonal variation — such as temperature differences, or changes in conditions underground — were not fully accounted for. DAMA scientists say they have addressed these issues, and argue that the dark-matter particle must be something more exotic.

Hopes of moving the debate forward rest on an upgraded detector called DAMA/LIBRA, from which results are expected before the end of 2008. **J.H.**

BENDING THE RULES

Careful measurements in the 1970s showed that the outer stars of galaxies travel too quickly for the gravity of the galaxies' visible mass to hold on to them. Dark matter was invoked as one way to provide the extra gravity needed. But not everyone agrees that a new form of matter is necessary. Some scientists proposed modified newtonian dynamics, or MOND, as a way to explain the observations.

First suggested in 1981, MOND holds that gravity weakens less slowly with distance than expected. The modern version of the theory has a few supporters, but it hasn't gained wider acceptance because there are many cosmic observations it struggles to explain.

MOND seemed to get a further knock last summer when NASA announced "direct proof of dark matter" from observations of two colliding galactic clusters.

NASA/M. MARKEVITCH *ET AL.*/D. CLOWE *ET AL.*

Astronomers can study dark matter by looking at how light has been bent around galaxy clusters, a phenomenon known as 'gravitational lensing'. Lensing revealed the mass distribution of the clusters to be distinct from that of the clusters' hot gas.

Researchers concluded that the 'bullet cluster' (pictured) had formed when one cluster tore through another. They suggested that the galaxies' gas got caught up in the collision, but the dark-matter particles shot straight through.

MOND researcher HongSheng Zhao of the University of St Andrews, UK, agrees that this is evidence for some kind of dark particle. But he doesn't accept that it rules out MOND — arguing that MOND plus ordinary neutrinos can explain the observations. Pursuing alternative theories is healthy, says Zhao, and this interpretation will be tested by future measurements of neutrino masses. **J.H.**

additions, but now some of them are leading candidates for dark matter. "These days a theory without a dark-matter candidate is not considered an interesting one," says Roszkowski. "The existence of the dark-matter problem is perhaps the most convincing evidence for physics beyond the standard model."

Many of today's leading theories for physics beyond the standard model are variations of 'supersymmetry', which posits that each ordinary particle has a heavier supersymmetric partner. Several of these partners have been put forward as candidate WIMPs, and, remarkably, calculations of the number of such WIMPs expected to be left over from the Big Bang match cosmological observations of dark matter. This coincidence helped to strengthen the case for dark matter being a new kind of particle, although the numbers take some getting used to. Assuming a typical WIMP has a mass 100 times heavier than a proton, models of the dark matter in the Milky Way predict there will be roughly ten billion WIMPs passing through one square metre of Earth every second. For these particles to zip by unnoticed requires ordinary matter and light to barely register their presence.

It also makes WIMPs incredibly difficult to spot. Calculations suggest that almost a million billion dark-matter particles pass through Aprile's XENON10 detector every week — yet only the tiniest fraction would ever be detected. The XENON10 experiment works on the principle that a passing WIMP should very occasionally bump into a xenon atom — a fat target, with 54 protons and 54 electrons. Such collisions would release energy through a handful of photons and electrons, which can be detected by sensitive instruments.

In common with other experiments that aim to directly detect dark matter, XENON10 is housed underground because the rocks above it absorb particles and radiation, such as cosmic rays from outer space, that might otherwise confuse the data. The challenge for the experimental teams is to block out as much of this 'background' as possible and see what's left.

Aprile announced XENON10's first results earlier this year[1]. The findings are yet to be published, but they took the community by surprise, not least because the previous best result belonged to an underground experiment using totally different technology. Rather than trying to trap dark-matter particles in a 15-kilogram vat of xenon liquid as the XENON10 detector does, the Cryogenic Dark Matter Search (CDMS) in Minnesota looks for vibrations and charge created by particle collisions in a very cold crystal of germanium and silicon. Until Aprile's result, experiments with noble liquids had been lagging behind.

In its first run, XENON10 registered 10 events over 60 days that couldn't be instantly dismissed. "You could jump up and down and say we've found ten WIMPs, but of course we haven't," she says. Aprile's team ruled out half of the events on closer inspection, and the rest were assumed to be background signals that slipped through the analysis. The researchers would have needed at least 15 events that could not be explained in other ways to think they'd caught a whiff of a WIMP, and Aprile says that, even then, they would need to understand the background better before claiming a direct detection.

Setting limits

In the meantime, a negative result is still important. The sensitivity of the XENON10 detector allows the researchers to set limits on the properties that a hypothetical WIMP might have — such as how heavy it is and how much it interacts with matter. This is crucial information when what you are hunting for is as mysterious as a dark-matter particle. XENON10 now claims a tighter limit than the previous best result. The experiments are starting to eat into the regions where supersymmetry predicts WIMPs should be (see 'WIMP hunting').

Other projects are close on XENON10's heels, searching for particles that might interact even more feebly. To improve their chances of finding a particle, dark-matter detectors need to do two things: get bigger and reduce the background.

"You could jump up and down and say we've found ten WIMPS, but of course we haven't."
— Elena Aprile

Designers of a rival xenon-based experiment, Zeplin-III, have taken great care to minimize stray signals reaching their detector, and this experiment is expected to yield results within the next year or two. The Zeplin project, a UK collaboration, started life in the early 1990s and the team had been in talks with Aprile before she decided to strike off on her own. "It's frustrating that they came into this game when we already had the Zeplin designs and picked off the best bits," says project spokesperson Tim Sumner of Imperial College, London, "but we should take the positive out of it; it shows the technology works."

With their latest 12-kilogram detector

installed in the corner of a potash mine in Cleveland, UK, the group is confident: "We have faith that Zeplin-III will be better than XENON10," says Sumner. This will help set even tighter limits on background noise, but "the glory will come with the first detection".

Closer to home, XENON10 has several rivals at Gran Sasso. The container that Aprile snapped a picture of during her tour will soon be the centrepiece of WARP, the WIMP Argon Programme, an experiment led by Carlo Rubbia, who won a share of the Nobel Prize in Physics in 1984 for his part in the discovery at CERN of the force-carrying particles known as the *W* and *Z* bosons. WARP will use a large vat of liquid argon to trap dark matter. "This is serious competition," says Aprile, who is a former student of Rubbia's.

Aprile is also upgrading XENON10 over the next few months by reducing the background and increasing the detector's size to 60 kilograms of liquid. A bigger vat boosts the chances of finding a WIMP, because having more mass makes it more likely that a dark-matter particle will interact. "The next step that is sensible is to go to one tonne; it doesn't make any sense any more to screw around with these little things," Aprile says.

Paying the price

But bigger detectors require more money. In a report published on 13 July, the Dark Matter Scientific Assessment Group, established last year by the US Department of Energy and National Science Foundation, recommended that US funding for dark-matter research be bumped up from less than $4 million per year to $10 million annually. This won't fund experiments at the tonne-scale, but it should accelerate testing of the different technologies, says Hank Sobel, chair of the group. The panel recommends that another review be carried out in 2009 to decide how to move to large-scale detectors.

It is easier to scale up the liquid approaches than the solid-state experiments, but the CDMS detector can more easily identify and eliminate background. The CDMS did undergo an upgrade a year ago, from one kilogram of germanium to four. That should mean that its next result, due at the end of the summer, will equal or better that of XENON10. "Over the next year, we should be able to increase our limits by at least a factor of 10 or 15, or discover something," says Sadoulet, who is the spokesperson for the CDMS.

Despite the enthusiasm, there is still a chance that nature will refuse to cooperate, and the experiments will chase ever better limits but never detect a particle. Some of the WIMP candidates predicted by supersymmetry are "essentially undetectable", warns Roberto Trotta, a theoretical physicist at the University of Oxford, UK. The particles may be too heavy to be created by the LHC and at the same time too weakly interacting to be detected by the underground experiments. "I think in the next ten years, we are going to see big discoveries; if not we are going to be in big trouble," says Trotta.

A. ROSZKOWSKA

"These days a theory without a dark-matter candidate is not considered an interesting one." — Leszek Roszkowski

When the LHC smashes together its protons in 2008, WIMPS might be created in the messy outpouring. The collider's detectors wouldn't be able to register these directly, but they would show up as 'missing mass' when the physicists piece together the energy budget of the collisions. Because such evidence is indirect, finding a WIMP signature at the LHC would not confirm it to be dark matter. "There would still be a window open," says Roszkowski. For example, a particle might be stable for the fractions of a second that it takes to fly out of the collider, but then decay elsewhere. That means a second route — direct detection — would be necessary too. "We absolutely need both to resolve the dark-matter problem," he says.

But the collider can provide additional theoretical context. For instance, should it be lucky enough to identify a particle, the LHC should be able to place it in a supersymmetric family tree.

Different supersymmetry theories predict superpartners with different masses — including the axino, gravitino or neutralino — as WIMP candidates. The dark-matter candidate is always the lightest partner, because this can't decay into anything else. Many of the simplest models predict that the neutralino, a particle that is a combination of the superpartners of four other particles, will be the lightest, and thus stable enough to have survived from the Big Bang until now. Consequently, this strange particle — it acts as its own antiparticle — is the most popular WIMP candidate among particle physicists.

A third route to detecting neutralinos will search for evidence of their destruction. Because

L. ROSZKOWSKI *ET AL. J. HIGH ENERGY PHYS.* (IN THE PRESS)

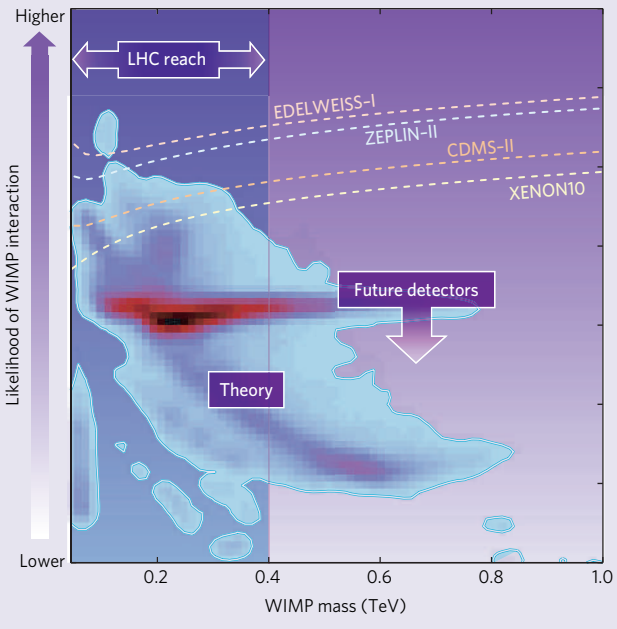
WIMP HUNTING

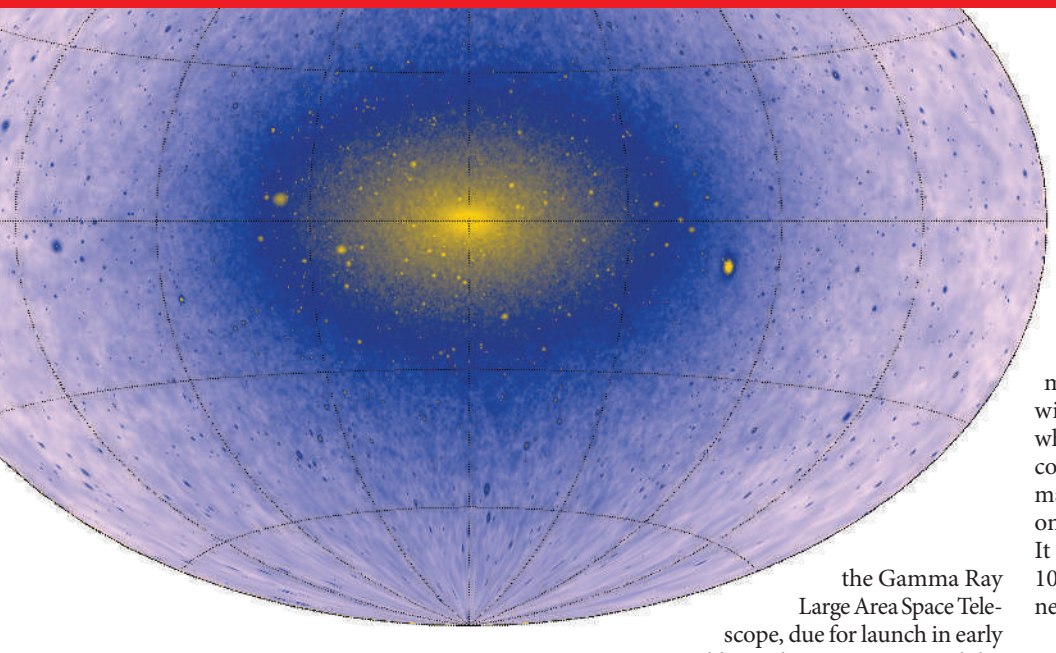
So far, detectors hoping to catch dark-matter particles known as WIMPS haven't found a thing, but it isn't all bad news.

Physics models that predict the existence of WIMPs can be used to calculate their likely properties. In this plot, the splodges define one guess about what the neutralino (a popular WIMP candidate) might look like[2].

Particles with properties above the detector-sensitivity limits (see lines across the plot) are ruled out by the experimental results so far; anything below them, including most of the theoretical splodge, is still possible.

Within the next few years, the detectors should reach the red heart of the splodge, where calculations say the neutralino is most likely to lie. Meanwhile, the Large Hadron Collider might make some neutralinos, but only if their mass is light enough. **J.H.**





M. KUHLEN, J. DIEMAND & P. MADAU

Now you see it: dark matter in the Milky Way could generate a γ-ray glow.

neutralinos are their own antiparticles they should annihilate each other in regions where they are close enough together to bump into each other. This self-destruction could show up in the form of γ-rays, neutrinos or matter and antimatter particles.

For example, models of dark matter in the Milky Way suggest that annihilation of neutralinos concentrated in the Galaxy's dark-matter halo and in the galactic core could generate a γ-ray glow (see picture). Satellites such as the Gamma Ray Large Area Space Telescope, due for launch in early 2008, could spot this. Moreover, annihilation of neutralinos clumped in the core of the Sun could be inferred from measurements made by neutrino telescopes, such as the IceCube Neutrino Detector being built at the South Pole, because such annihilation would generate higher-energy neutrinos than expected from other processes.

But what if dark matter isn't a neutralino or even a WIMP? Some proposals to explain dark matter don't depend on supersymmetry or WIMPS at all. These range from doing away with the need for dark matter by modifying the laws of gravity (see 'Bending the rules') to suggesting other types of particle altogether.

The main rival to the neutralino is the axion, first proposed by particle physicists in 1977 to resolve a glitch in the standard model. Many theorists believe that the axion will eventually be found, but it is unclear whether its mass and interactions will match cosmological expectations. Already the axion mass is constrained on one end by theory and on the other by observations of supernovae. It is consequently predicted to be at least 10 million million times lighter than a typical neutralino.

The Axion Dark Matter Experiment, at the Lawrence Livermore National Laboratory in California, aims to give a definitive answer to the axions' part in dark matter. Leslie Rosenberg, co-spokesperson for the experiment, says the team expects results by 2011, after the detector moves to its final home at the University of Washington in Seattle. Axions interact so weakly that they are rarely produced in particle colliders, so the experiment will look for signs of axions using a radio receiver that can detect tiny particle energies.

There are far fewer searches for axions than there are searches for WIMPs. Rosenberg thinks this is partly because the technology

THE SEARCH FOR STRUCTURE

When it was first conceived, almost a decade ago, it was known as the Dark Matter Telescope. With one of the largest astronomical mirrors ever cast, and a unique wide field of view, it was designed to pick up the faint lensing of light produced by clumps of dark matter from distant galaxies, revealing dark matter's mysterious behaviour and, perhaps, its nature.

It has been renamed the Large Synoptic Survey Telescope (LSST). The more generic name reflects the degree to which the telescope's capabilities will be exploited by an ever wider range of astronomers. When the LSST starts its surveys sometime next decade, dark-matter mavens peering into the depths of space will rub virtual shoulders with those looking for the asteroids closest to Earth — and with the aficionados of dark energy. Although dark energy had not even been discovered when the LSST was first dreamt up, the new telescope should provide the sort of structural data that scientists need to deepen their understanding of the Universe's acceleration (see 'A constant problem', opposite).

But despite the LSST's accumulated endorsements, it has not reached the front of the queue for receiving money from the US National Science Foundation. Without a firm commitment on its $375-million budget, no one can say for sure when construction at the top of Cerro Pachón, the 2,682-metre peak in Chile that has been chosen as the site, will actually get under way. Meanwhile, an upstart rival, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) is already being assembled atop Mount Haleakala in Hawaii with the help of funding extracted from the Department of Defense by a friendly senator.

Thanks to Senator Daniel Inouye (Democrat, Hawaii) Pan-STARRS gets about $10 million a year through the Air Force as a congressional 'earmark', and the team involved is using that cash to get its telescope up and running as soon as possible. Team members have bought electronic equipment off eBay to save time and money, and the telescope design is simple and quickly assembled, according to project director Kenneth Chambers, an astronomer at the University of Hawaii in Honolulu.

The first of the 1.8-metre telescopes built to the Pan-STARRS design, called PS1 (pictured opposite), saw first light in June of last year. The team is currently installing a 1.4-gigapixel camera — ten times the size of the camera used for the Sloan Digital Sky Survey, the most accomplished deep-space survey to date — and hopes to have its first survey data by autumn. The novel part of the project will come later, though, when this first telescope gets three identical companions.

The idea of multiple mirrors and cameras is not to increase resolution, as an interferometer does, but to improve sensitivity. All the signals from a given object will be added together by computers no matter which camera recorded them. The distributed design also provides the system with effective immunity to false data created by cosmic rays striking individual cameras, says Nick Kaiser, Pan-STARRS principal investigator and an astronomer at the University of Hawaii in Honolulu.

Group members are bullish about the possibility of getting significant findings on dark matter and energy before the LSST comes online. "The real advantage of Pan-STARRS over the LSST is that being a small, fast, somewhat entrepreneurial group, we're there

needed to find them is less familiar to particle physicists. In addition, he says, "there's a tremendous bandwagon to supersymmetry, and WIMPS are riding on that". Even if dark matter turns out to be something completely different, the experimental teams are determined to track down their particular quarry and get an answer, one way or another. "I live with that with the impatience of the Italian woman that I am," says Aprile. "I am just going fast ahead with the next step, making the detector better."

But success for one type of experiment doesn't have to mean failure for another. Scientists prefer simplicity: if they find WIMPS, then they don't need axions, and vice versa. But why not have both? Dark matter might prove to be a richer problem than anyone is expecting. Tegmark hopes for this outcome. "This could be a wonderful surprise. It's very arrogant of us humans to say that just because we can't see it, there's only one kind of dark matter." ■

Jenny Hogan is a reporter for *Nature* in London.

1. Angle, J. *et al.* preprint at http://arxiv.org/abs/0706.0039 (2007).
2. Roszkowski, L., Ruiz de Austri, R. & Trotta, R. preprint at http://arxiv.org/abs/0705.2012 (2007).

See Insight, page 269, and Editorial, page 225.

A constant problem

Why is dark energy, hailed as a breakthrough when discovered a decade ago, proving so frustrating to the scientists who study it?

In 1998, two teams of astronomers reported that the Universe was pulling itself apart. This came as something of a shock. That the Universe was expanding had been known since the 1920s, but conventional wisdom held that this expansion was slowing and was likely, in the distant future, to come to an all but complete halt. Then, in the late 1990s, observations of distant supernovae showed that the expansion was not slowing down at all. It was speeding up. This discovery was incredibly counterintuitive, recalls Charles Bennett, an astronomer at Johns Hopkins University in Baltimore, Maryland. "I just didn't believe it."

Within a few years, however, he and almost all his peers could withhold their belief no longer. The observations became stronger. And the expansion provided a way out of a theoretical impasse. Observations of the Big Bang's afterglow made by various groups, including Bennett's, indicated that the Universe's gravity had flattened it out. But other observations suggested that it simply didn't contain enough matter to have that much of a gravitational effect — even when as-yet-undiscovered forms of dark matter were included in the sums (see page 240).

Happily, the theory of relativity requires energy, as well as matter, to have a gravitational effect. And it turned out that the amount of energy needed to drive the acceleration was pretty close to that needed to solve the flatness problem by means of its gravity. 'Dark energy', as it quickly became known, seemed poised to provide great insight into the origin and future of the cosmos, says Michael Turner, a cosmologist at the University of Chicago in Illinois. "This seemed to be the piece that made everything else work."

But a decade further on, researchers seem to have swapped one theoretical conundrum for a bigger one. Follow-up measurements have

seven or eight years earlier," says Chambers. But integrating the four data sets and processing the data will be no mean feat; building the first telescope was, comparatively speaking, the easy part.

The LSST is revolutionary in different ways. On top of its unique mirror design, there's the sheer amount of data that it will accumulate. Its 3.2-gigapixel camera should, over the course of the telescope's life, produce more than 100 petabytes of data. That's as much information as contained in the whole genome of every animal on Earth, according to Tony Tyson, the astronomer at the University of California, Davis, who has headed the LSST project since its days as the Dark Matter Telescope. The sheer amount of data to be made sense of is one of the reasons the LSST is happy to have formed a partnership with Google. The Sloan Survey gathered data at a rate of 200 gigabytes a night; the LSST is aiming for 30 terabytes.

B. SIMISON

The databases produced by both the LSST and Pan-STARRS will provide astronomers with more than just measurements of dark energy and matter. Both telescopes plan to image wide swaths of the sky multiple times, allowing astronomers to spot things moving in the Solar System, as well as changing phenomena in the depths of the sky. The potential for discovery is enormous, says Tyson. Kaiser agrees. "You're going to get a sort of movie of the sky," he says.

For now, it seems that Pan-STARRS has the edge in the race to map out the Universe's darkest quarters. But if the LSST team is put out, then the group does its best not to show it. "If they make discoveries before LSST gets online, great," says Steven Khan, the LSST deputy director at the Stanford Linear Accelerator Center in California. "To date it hasn't really been a problem." "It's healthy to have both Pan-STARRS and LSST," Tyson adds. **G.B.**

NASA; J. BLAKESLEE, JOHNS HOPKINS UNIV.)

Successive images of a patch of sky can reveal exploding supernovae (red spot).

revealed little about the nature of dark energy, and theories to explain it have failed to gain traction. And although astronomers are trudging forwards with a battery of new measurements, there is little guarantee that any will solve the problem — and thus no clear consensus on how much effort to put into them. "The issue is: how much information do we get from these future observations?" asks Avi Loeb, an astrophysicist at Harvard University.

Hidden depths

The big problem is that dark energy is not, in itself, something that astronomers can see. Like dark matter, it is known only by its effects — in this case, the effect it has on the Universe's acceleration. The acceleration is related to dark energy through a quantity known as the 'equation of state'— the ratio of the pressure dark energy exerts to the energy per unit volume involved.

An accelerating expansion means that the equation of state has to be negative. And a value of −1 would mean that dark energy was an unchanging feature of the cosmos — a 'cosmological constant'. Such a constant had been a feature of Einstein's general theory of relativity, one that he had added, ironically, as a way of guaranteeing that the Universe would stay the same size. When Einstein came to accept that the Universe was, in fact, expanding he removed the term, calling it his "greatest mistake". But if the equation of state had a value of −1, dark energy would fit the cosmological constant bill perfectly. And current measurements make it quite possible that the equation's value is −1.

If dark energy's equation of state is indeed −1, then there's one obvious way to make sense of it, says Leonard Susskind, a cosmologist at Stanford University in California. For decades, physicists have postulated the existence of something known as 'vacuum energy' — a primordial froth of quantum particles that flit in and out of existence in the vacuum of space. This vacuum energy could drive the observed accelerating expansion, and it would do so in a constant manner. Because vacuum energy is an inherent property of space, Susskind explains, an expanding Universe would create more of it, meaning that the ratio of energy density to pressure would never change, their ratio fixed for ever at −1.

There's just one theoretical discrepancy: the vacuum energy as calculated by physicists is more than 10^{100} times larger than would be needed to explain the relatively weak effects of dark energy as observed by astronomers. If it were as big as physicists suggest, then our Universe would fly apart in the blink of an eye. "Every calculation indicates that vacuum energy should be enormous," says Turner. "There's no natural way to get such a tiny number." So most physicists have hoped that some yet-to-be-discovered effect based on some hidden symmetry of nature would cancel out the vacuum energy. Such a hope-it-goes-away approach is used by physicists quite a lot, and can be the only way to make progress in some circumstances. At the same time, applying it to the vacuum energy was, admits Susskind, "completely illogical".

"And I must say I shared that illogical attitude myself," he continues almost apologetically. Now, he thinks differently, and is one of those who has proposed a solution of sorts to the conundrum. 'String theories', popular with many particle physicists, make it possible, even desirable, to think that the observable Universe is just one of 10^{500} universes in a grander 'multiverse', says Susskind. The vacuum energy will have different values in different universes, and in many or most it might indeed be vast. But it must be small in ours because it is only in such a universe that observers such as ourselves can evolve.

This sort of anthropic argument irks many scientists. Critics say such reasoning is almost impossible to verify and doesn't provide any deeper insight into the cosmos. "Anthropics and randomness don't explain anything," says Paul Steinhardt, a theorist at Princeton University in New Jersey. "I'm disappointed with what most theorists are willing to accept."

The trouble is that no

other approaches are proving any more fruitful. Some suggest that the problem lies with Einstein's idea of gravity, which they then seek to modify in a way that fits in with dark energy. "It would be very fortunate if the dark energy were a modification of gravity," says Georgi Dvali of New York University, "because it would address fundamental questions of physics." But others see little mileage in such changes. Leaving aside the cosmos, "it's not so easy to get those theories to be consistent with our Solar System", says Turner.

Another possibility is that dark energy is some sort of evolving property of the Universe. Some postulate that dark energy is a fifth force (the others being electromagnetism, the two nuclear forces and gravity) that works at the largest scales of the cosmos. Others suspect that it is the aftermath of the inflation that many see following directly on from the Big Bang. Inflation was, after all, a period of extreme expansion — might it not have some sort of 'long tail' that stretched away down cosmic history? These solutions and others, although different conceptually, are equivalent mathematically. And they share a requirement that dark energy changes over time — that its equation of state is not locked in as −1. Such a change would help to explain why dark energy is apparently so weak today, says Steinhardt. And changing values for dark energy might affect other features of the Universe, including some parameters now seen as fundamental constants, in detectable ways, which could be a plus. But critics claim that these ideas require extreme amounts of special pleading.

D. PARKINS

Starring role

In general, the theoretical side of the debate is not a pretty thing. "We've tried a whole bunch of things and nothing has sprung forward," says Sean Carroll, a theoretical physicist at the California Institute of Technology in Pasadena. What's needed, Carroll says, are a few more good clues.

Astronomers are planning a new generation of dark-energy probes that will refine measurements of the equation of state. They are already pushing ahead with further measurements of type 1a supernovae. These stellar outbursts occur when a stream of material being sucked from a larger star onto a smaller one pushes the smaller star's mass over a threshold, precipitating a massive thermonuclear explosion. Because each star explodes at the same mass threshold, they should all give off the same amount of energy. And so, in absolute terms, each should be as bright as any other. By comparing their relative brightnesses when seen from Earth, it is possible to measure the distance to the explosion with precision, says Saul Perlmutter, the astronomer at Lawrence Berkeley National Laboratory in California who led one of the original dark-energy supernova teams. And by measuring distance in this way and speed by means of the 'red shift' of the supernova's light, astronomers can understand acceleration over time. Perlmutter and others are now working to increase both their understanding of the supernova mechanism and the size of their sample to improve on their original calculations.

A. WARREN

"We could be deeply wrong about cosmology for the next thousand years."
— Leonard Susskind

Supernovae, although the best understood, are not the only way to measure acceleration. Another option is to study X-rays from distant clusters of galaxies. As in the case of supernovae, a cluster's temperature and brightness should have a standard relationship, so it should be possible to measure the speed at which those at a given distance from Earth are receding, says Steve Allen, an X-ray astronomer at Stanford University.

It is also possible to measure the effects of dark energy in subtler ways. The gravitational field of a cluster or group of galaxies makes light shift towards the blue as it falls into the galaxies' gravity well, and reddens it as it climbs back out. According to Ryan Scranton, an astronomer at the University of Pittsburgh in Pennsylvania, dark energy should affect the way these effects show up in the cosmic microwave background, radiation left over from the Big Bang.

A tangled web

Combining these different sorts of measurement should offer ways of constraining the value of the equation of state better than any single measurement can manage (see 'Closing in on dark energy', overleaf). Perhaps the most promising new realm of research, say many in the field, lies in surveys that will look at how the largest structures in the Universe have been shaped or distorted by dark energy. Galaxies are not spread evenly across the cosmos, but instead clump into a three-dimensional cobweb. The structure of that cobweb is sensitive to dark energy. And the sort of error to be expected in measurements of the structure are completely different from those that plague measurements of supernovae, according to Adam Riess, an astronomer at Johns Hopkins University who led the original supernova team that competed with Perlmutter's. That makes the new approach pleasingly independent of the old one. Several ambitious surveys are now being planned to further map the large-scale structure of the Universe (see 'The search for structure', page 244).

But none of these techniques can do more than narrow the frustratingly uninformative

equation of state down further. To prove that dark energy is a cosmological constant requires showing that the equation of state is indeed −1. Merely showing that it is close doesn't cut it. Astronomers could basically go on measuring dark energy for ever without eliminating other possible theories, says Simon White, director of the Max Planck Institute for Astrophysics in Garching, Germany: "If it's just a constant, then you need infinite accuracy."

Lawrence Krauss, a theoretical physicist at Case Western Reserve University in Cleveland, Ohio, goes further. If the equation of state is indeed −1, and dark energy is a constant, then the only way to measure it will be through its effect on the Universe's acceleration. "If it is −1, we won't know what dark energy is," he says. "It doesn't give us any theoretical guidance whatsoever."

Carlos Frenk, a theoretical physicist at the University of Durham, UK, agrees that probing a single number without a strong theoretical case for doing so is not the way forward. "It's like trying to learn something fundamental about biology by measuring the height of every tree," he says. "Just measuring something for the sake of measuring it is pointless." Frenk questions how much money should be spent on such measurements, and Loeb agrees. "One should put money in this direction," he says, "but not excessive amounts."

But for all the worries of some theoretical physicists, observational astronomers think that carrying on with the equation of state measurement is the most sensible next step, not least because it is the only one on offer. "My feeling is that we should measure it to the limits," says Bennett. "We may see things that surprise people, that often happens." Perlmutter, too, sees room for a few more results to narrow things down rather than shaking them up. "It seems like you'd want to get a couple of boring results before you decide 'we're done'," he says.

Bennett and Perlmutter's enthusiasm for further measurements is evidenced by the fact that they are heading up rival proposals for spacecraft to observe galactic structure (Bennett) and distant supernovae (Perlmutter), seeking to get the money that NASA and the US Department of Energy are considering spending on a dark-energy probe. And even if their endeavours contribute no more than some incremental precision to the debate on dark energy, the observations will still tell astronomers quite a bit about other things in the Universe. A space-based supernova probe, for example, would provide a high-quality survey of infrared objects throughout the sky. "These are not special-purpose instruments," says Roger Blandford, director of the Kavli Institute for Particle Astrophysics and Cosmology in Stanford, California. "They will revolutionize a whole range of fields."

UNIV. CHICAGO

"Dark energy seemed to be the piece that made everything else work."
— Michael Turner

And there is always a chance that some other area will reveal the next much-needed clues as to the nature of dark energy. When it starts taking data in 2008, the Large Hadron Collider at CERN, the particle-physics laboratory near Geneva, might conceivably make relevant discoveries about the nature of space-time (see Insight, page 269); "We may learn more from accelerators than we do from the sky," says Krauss. Similarly, measurements of fundamental constants and gravitation at short distances could have some unexpected connection to the dark-energy problem; and detecting some sort of dark matter might help, too (see 'Welcome to the dark side', page 240).

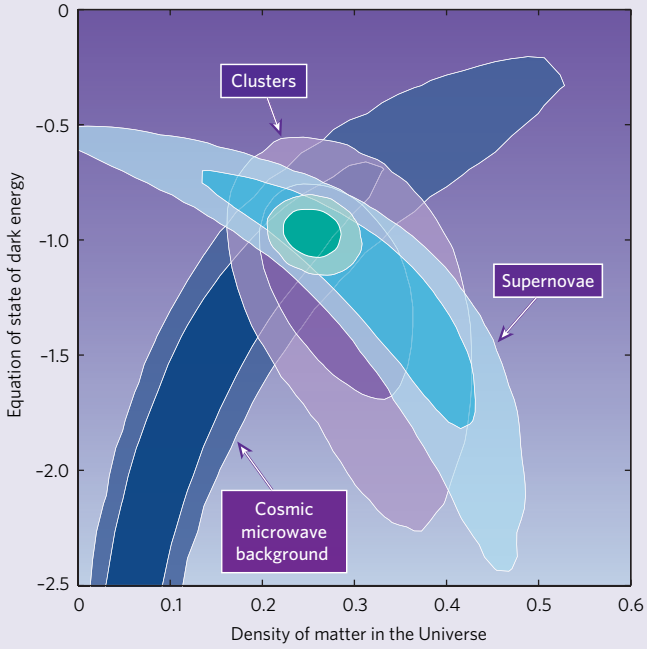
So far, though, the revolution promised by dark energy's discovery a decade ago hasn't materialized. Although researchers are more certain than ever of the existence of a cosmic push, they know as little about what it means physically as they did in 1998. "Right now there are two possibilities," says Carroll. "Dark energy is vacuum energy, or it's something else." Observers are slightly more upbeat. "It feels to me like a very early discussion of all this," says Perlmutter. Still, he concedes, without a measurement of the equation of state that deviates from −1 it will be difficult to learn much of anything. "If you don't see those ripples," he says, "it's going to be hard to play the game."

For now, many in the field are left with a sense of unease: the tantalizing clue they thought they had discovered has turned into an exasperating mystery. And with no clear explanation of something that could be up to three-quarters of everything out there, it's hard not to feel like you're missing a big part of the picture, Susskind says. "We could be wrong about cosmology for the next thousand years. Deeply wrong." ■

Geoff Brumfiel is *Nature*'s physical sciences reporter in Washington DC.

See Editorial, page 225

CLOSING IN ON DARK ENERGY



Various observations constrain the possibilities for the Universe's density and the dark-energy equation of state in coordinated ways. Measurements of the cosmic microwave background, which show that the Universe is flat, rule out all values for the equation of state and the density of matter outside the purple banana shape from bottom left to top right. Measurements of supernovae, which demonstrate acceleration, rule out values outside the pale and darker blue banana that goes from bottom right to top left. Measurements of galaxy clusters, another approach to acceleration, rule out everywhere not in the pink/mauve region. A statistical combination of all the measurements gives the green spot. In all cases the contours define regions ruled out to 68% certainty (inner contour) and 95% certainty (outer contour).

S. W. ALLEN *ET AL.* PREPRINT AT HTTP://ARXIV.ORG/ABS/0706.0033 (2007)

Race: talented minorities face a 'revolving door'

SIR — Your News story 'Researcher refuses to back down over race case' (*Nature* **447,** 762–763; 2007) calls attention to the courageous stand taken by James Sherley, an assistant professor at Massachusetts Institute of Technology (MIT) who believes that he was denied tenure because of racial discrimination.

Remarkably, although there has long been a high percentage of African-American students at leading US universities — 10% of incoming undergraduates at MIT, for example — very few have so far made it through to tenure. Only about 1% of biology professors at US universities are African-Americans. Although one-third of assistant professors overall may make it to tenure at MIT, hardly any African-American assistant professors have ever done so in MIT's core disciplines of science and engineering. The same seems to be true at most leading US universities, leading to what has been termed a 'revolving door' for even very talented young African-American scientists such as Sherley, who last year won a National Institutes of Health Pioneer award for innovative work.

How can it be that white and black scientists who initially seem equally talented have such different chances of making it to tenure? I would argue that it is because present tenure policies are unintentionally designed to prevent the success of even the most talented minority scientists.

At places like MIT, only a fraction of faculty make it, even if they're white. In the face of pervasive racial barriers, how can talented minorities have a fair chance in such a steeply competitive timed-tenure system? These barriers can include lack of equal space and resources, lack of mentoring by senior faculty, lack of inclusion in faculty activities such as invitations to speak in seminar series, a general lack of recognition and support, and a hesitancy among white students to join the labs of minority faculty or to be referred to minority labs by senior faculty.

In the face of so many obstacles, how is it fair to argue that Sherley does not deserve tenure because he didn't publish quite as many papers as white assistant professors who did not face any of these barriers? Although the MIT faculty and administrators who have considered his tenure application are for the most part well-meaning, they seem to be unaware of the reality of persisting racial barriers. They unfairly prefer to attribute lack of success to inability.

In a survey of MIT students in 1985, it was found that African-Americans have surprisingly few meaningful faculty contacts, most of those being with the tiny percentage of the faculty who are ethnic minorities. If minority students and faculty are to be successful, there is an urgent need for universities to re-evaluate and redesign their policies that control retention of ethnic minorities on their faculty.

Ben Barres
Department of Neurobiology,
Stanford University School of Medicine,
Fairchild Room D235, 299 Campus Drive,
Stanford, California 94305-5125, USA

Race and tenure case was not handled fairly by MIT

SIR — Although tenure evaluations are not primarily accountings of publications, you reported in your News story 'Researcher refuses to back down over race case' (*Nature* **447,** 762–763; 2007) that I published six peer-reviewed research papers during the years before the decision taken by Massachusetts Institute of Technology (MIT) about my tenure.

My years as a principal investigator before MIT's decision include research at the Fox Chase Cancer Center in Philadelphia. MIT's tenure decision should have been based on my comprehensive work as a principal investigator, not limited only to time at MIT. The MIT faculty personnel record submitted for my tenure evaluation listed 41 scholarly articles published, in press, or accepted for publication, including 11 peer-reviewed primary-research articles, two peer-reviewed review articles, five peer-reviewed proceedings papers and four book chapters (two peer-reviewed). Not included in this total are four research manuscripts submitted to peer-reviewed journals and 10 published patent applications.

Your comparison of my tenure application with those of two other faculty awarded tenure at the same time is not a fair comparison, because people who arrive at an institution mid-career are not comparable to those who began their faculty careers at the institution at which they later apply for tenure. Their research programmes are at a different stage of maturity, and often the projects undertaken differ significantly in degree of challenge and impact. Even so, another mid-career faculty member received MIT tenure within the same timeframe as my application, largely on the basis of contributions that had been made before arrival there.

My main complaint against MIT is the manner in which my case was decided by the faculty chair. For example, at MIT, when a tenure-case decision is being made, review of the case is prohibited outside its department. If the case is not advanced to the next level of review, it is sealed. So why was a professor who is neither a member of my faculty nor an expert in my field — stem-cell biology — asked by the faculty chair to review the case before the decision was announced?

James Sherley
Department of Biological Engineering,
Biotechnology Process Engineering Center,
Center for Environmental Health Sciences,
Center for Cancer Research,
MIT, Cambridge, Massachusetts 02139, USA

Foundation active in fight to cure Huntington's

SIR — Your News Feature on biomedical philanthropy 'Love or money' (*Nature* **447,** 252–253; 2007) states that that the High Q Foundation is a successor to the Hereditary Disease Foundation (www.hdfoundation.org). You also state that CHDI is a successor to the Cure Huntington's Disease Initiative of the Hereditary Disease Foundation. The word 'successor' could give readers the false impression that the Hereditary Disease Foundation either no longer exists or is no longer active. This is incorrect. The Hereditary Disease Foundation, since its inception in 1968, has vigorously encouraged and respectfully supported researchers seeking to find treatments and cures for Huntington's disease as rapidly as possible and continues to do so.

Nancy Wexler*†, Carl Johnson†
*Departments of Neurology and Psychiatry, Columbia University, 1051 Riverside Drive, Unit 6, PI Annex 371, New York, New York 10032, USA
†Hereditary Disease Foundation, 3960 Broadway, 6th Floor, New York, New York 10032, USA

Friendly clarification from City of Brotherly Love

SIR — I thoroughly enjoy *Nature*'s insightful columns and News and Views, and of course take a special satisfaction when researchers from my institution, the University of Pennsylvania (or U Penn) are featured. However, in a sidebar 'Body and mind' within the News Feature 'Brain craze' (*Nature* **447,** 18–20; 2007), a researcher is reported to work at Pennsylvania State University (Penn State) in Philadelphia. There have been occasional other instances in your pages of confusion over Pennsylvania and its universities.

Neither Penn State nor the University of Pittsburgh (Pitt) is in Philadelphia, the state's largest city, located in its southeastern corner. U Penn is a private university in the city, founded by its most famous citizen, Benjamin Franklin.

Penn State is the state's (and maybe the country's) largest public university, and it has many campuses, the main one located in University Park, Pennsylvania. The legend is that the location of the university was chosen

by drawing a large 'X' from the four corners of the state and placing the university at the centre of it — in order to make it equally accessible to all students.

The University of Pittsburgh (Pitt) is a public university that resides, as one might have guessed, in Pittsburgh, the second largest city in Pennsylvania and at the western edge of the state.

Douglas J. Jerolmack
Department of Earth and Environmental Science, University of Pennsylvania, Hayden Hall, 240 South 33rd Street, Philadelphia, Pennsylvania 19104, USA

Animal welfare is not just another bureaucratic hoop

SIR — I disagree with C. Jimenez's reply, in Correspondence, opposing Victoria Buck's suggestion of making animal-welfare sections in scientific papers compulsory ('Animal-welfare section in papers would be a burden' *Nature* **447,** 259; 2007). We all have a great many bureaucratic hoops to jump through these days, but we should not take a dismissive attitude to animal-welfare issues.

Animal-rights extremists have made life a misery for some scientists in the United Kingdom, despite our having one of the best-regulated licensing and ethical review processes in the world. National legislation requires scientists wishing to carry out experiments on animals to be licensed, and strict enforcement by both the legislature and by the local ethical review committees ensures that there are very few infringements.

The exchange between Buck and Jimenez did not address ethical approval statements, but for the record I do not think it an onerous task to include in scientific papers a paragraph stating the legislation(s) and local ethical review process under which the work had been approved. Many journals, including the Nature journals, already make compliance a condition of publication (see www.nature.com/authors/editorial_policies/experimental.html).

Although we must be robust in our defence of the need for appropriate animal experimentation, it is pointless to antagonize those individuals who will never be persuaded of its need or relevance. The 3Rs requirement goes some way to assuaging the disquiet of the more reasonable objectors, and hence should not be dismissed.

We live in a cynical world where everything is questioned, and the scientist is no longer seen as an ivory-tower figure. We are all accountable to the agencies that fund us and regulate our use of experimental animals and human tissue samples. Our ability to pursue science gives us a privilege that few others enjoy, that of unravelling the biological processes that make us what we are. We are enabled in this occupation by the silent consensus, and hope, of people all over the world. We abuse that consensus at our peril.

L. Bergmeier
University of London, London, UK

Animal welfare: reporting details is good science

SIR — C. Jimenez, in Correspondence, considers that detailed information on the way animals are handled and treated should not be placed in published papers (*Nature* **447,** 259; 2007).

I disagree, because it is a fundamental principle of the scientific process that when a paper is published, the study can be repeated from the description given in the methods, thereby allowing external validity to be assessed. To this end, variables that might affect the results need to be reported accurately.

It is well-documented that making even a slight change to a laboratory animal's environment or husbandry can have profound influences on its biological functioning. Cage size can influence metabolism, baseline rectal temperature, the fever response, feeding behaviour and behavioural responses in predator–prey interactions. The type of flooring in a cage can affect blood pressure, heart rate and body temperature. Other factors that influence physiology and behaviour include housing laboratory mice as singletons or pairs, the complexity of the cage and the extent to which animals are handled.

Variables such as these, which might be changed to improve the welfare of the animals, should be reported in published papers as an essential component of the accurate reporting of science.

C. M. Sherwin
University of Bristol, Bristol, UK

UNAIDS rejects claims of exaggeration and bias

SIR — We would like to provide our perspective on your Book Review of two books criticizing the Joint United Nations Programme on AIDS (UNAIDS), 'Time for a change?' (*Nature* **447,** 531–532; 2007), and the coverage of this issue at www.nature.com/news/2007/070528/full/070528-6.html.

In his book *The AIDS Pandemic: The Collision of Epidemiology with Political Correctness*, James Chin accuses UNAIDS of exaggerating data for the sake of advocacy, which is not true. Nor are UNAIDS data influenced by political or fundraising agendas. The UNAIDS Secretariat and the World Health Organization work closely with other technical partner organizations to assist countries in better understanding their HIV epidemics so they can respond appropriately. Estimations are produced in close collaboration with national epidemiologists and governments, using methodologies recommended by an international team of experts chaired by a leading academic from Imperial College London.

UNAIDS is committed to providing the most accurate information available and continues to be transparent in publicizing the methods used to assess the magnitude of the past and current epidemics. UNAIDS has always stated that countries should use the most comprehensive and most recent data available. Reassessments of earlier published estimates of prevalence, incidence and mortality have been made, and we expect that there may be adjustments in the future.

Helen Epstein's *The Invisible Cure* also makes inaccurate statements about the work of UNAIDS: in particular, we have always advocated the reduction of number of sexual partners as an effective strategy for HIV prevention, as can be seen from our reports and other contributions to the published record. All UNAIDS documents on the prevention of sexual transmission of HIV advocate abstinence, reduction of sexual partners and correct use of male and/or female condoms. (See, for example, http://data.unaids.org/Global-Reports/Bangkok-2004/UNAIDS_Bangkok_press/GAR2004_pdf/GAR2004_ExecSumm_en.pdf)

UNAIDS and its partners will continue their mission to gather the best-quality data to assist in shaping an effective global response to AIDS.

Paul R. De Lay*, Kevin M. De Cock†
*Evidence, Monitoring and Policy, UNAIDS, 20 Avenue Appia, 1211 Geneva, Switzerland
†HIV Department, World Health Organization, 19 Avenue Appia, 1211 Geneva, Switzerland

Chinese recorded classical nova two millennia ago

SIR — I agree with Michael M. Shara and colleagues (*Nature* **446,** 159–162; 2007) that the star Z Camelopardalis was a classical nova a few thousand years ago. In fact, a record of the eruption exists in Chinese documents of the time. There was, apparently, a report of a 'guest star' in October–November 77 BC (P. Y. Ho *Vistas Astron.* **5,** 127–225; 1962). The position in the sky fits Z Camelopardalis. This seems to be the oldest classical nova recorded in any surviving text.

Göran H. I. Johansson
Tordönsvägen 4G, 1tr, SE-22227 Lund, Sweden

BOOKS & ARTS

The case of creation

Last year's Dover trial resulted in intelligent design being removed from the science curriculum.

The Battle Over the Meaning of Everything: Evolution, Intelligent Design, and a School Board in Dover, PA
by Gordy Slack
Jossey-Bass: 2007. 240 pp. $24.95

40 Days and 40 Nights: Darwin, Intelligent Design, God, OxyContin® and Other Oddities on Trial in Pennsylvania
by Matthew Chapman
HarperCollins: 2007. 288 pp. $25.95

Monkey Girl: Evolution, Education, Religion, and the Battle for America's Soul
by Edward Humes
Ecco: 2007. 400 pp. $25.95

C. KASTER/AP PHOTO

Pennsylvania parents, and their children, fought against the teaching of intelligent design in schools.

Kevin Padian

Three new books use as a centrepiece the court case of Kitzmiller *et al*. versus Dover Area School District, which played out for six weeks in late 2005 at the state capital of Pennsylvania. This trial was the latest in a series of American 'Scopes trials', named after the 1925 prosecution of Tennessee teacher John Scopes, who was fined $100 for flouting a state law that prohibited the teaching of evolution in state-run schools. Scopes volunteered to be the test case, knowingly breaking the law. Famed attorneys Clarence Darrow and William Jennings Bryan argued the case. Scopes lost, Tennessee was ridiculed, a few other states passed similar legislation, and the divide between fundamentalists and secularists in the United States was irrevocably cleft.

Since the Scopes case, American jurisprudence has increasingly sided with the Enlightenment in a sequence of landmark decisions: yes, you can teach evolution; no, you cannot balance it with creationism; no, 'creation science' is not science; and so on. Then, in the late 1990s, a new kid on the block, intelligent design, began to flex its muscles and demand consideration as a viable scientific theory — but in the public arena, not the scientific one. Intelligent-design proponents, mostly right-wing Christians with more chutzpah than scientific acumen, gathered steam, money and eventually a grand strategy to "reverse the stifling dominance of the materialist world view, and to replace it with a science consonant with Christian and theistic convictions". They set up a think-tank, the Discovery Institute, and started to write op-ed pieces and lobby school districts to introduce their exciting new concept to children. A gullible and obstinate school board in the middle of Pennsylvania's rolling hills was just crazy enough to buy it — and that was the start of the now-famous Dover case.

The characters on all sides of the Dover trial — judge, plaintiffs, witnesses, school-board members and attorneys — are colourful and complex, and the trial strikes at the heart of what still divides the US population, 400 years after European settlers arrived. Is the American tradition one of philosophical and political idealists, or of persecuted pilgrims who then turn around and ostracize anyone who doesn't agree with them?

The three books take different tacks and each has different strengths. The author of *40 Days and 40 Nights*, Matthew Chapman, is a great-great-grandson of Charles Darwin; his presumed vested interest in the proceedings is tempered by his own history as a school dropout, a movie screenwriter and a Brit with a perpetually bemused view of colonial antics. Still, his odyssey is a fulfilling one, and he seems genuine enough to get himself invited into many homes where insights and passions run deep. Gordy Slack, author of *The Battle Over the Meaning of Everything* and an experienced science writer and editor, likewise brings his own family baggage (his father is a staunch fundamentalist) to his account, but his reporting is more linear and his background research deeper. Edward Humes in *Monkey Girl* is even more scholarly and thorough in his approach, and contextualizes the trial historically. Unlike Chapman and Slack, he does not insert himself into his narrative, but his views of the proceedings are no less clear.

The particulars of the trial are by now familiar (see *Nature* **437,** 607; 2005 and *Nature* **439,** 6–7; 2006). Dover's school board, against the advice of its teachers and attorney, required that high-school biology students be read a statement that among other things alleged that "gaps in the [evolutionary] theory exist for which there is no evidence" and that intelligent design is "an explanation of the origin of life that differs from Darwin's view". Students were referred to a supplementary text published by the Foundation for Thought and Ethics, *Of Pandas and People*, for more information on intelligent design. Eleven parents sued, engaging the American Civil Liberties Union and other top representation and scientific advice. The decision of Judge John E. Jones III slammed the "breathtaking inanity" of the school board, established its religious motive and actions, accepted the view of the scientific community that intelligent design does not qualify as science, and proscribed bogus criticisms of evolution in science classes. Intelligent-design proponents sputtered and

fumed; the usual right-wing commentators fulminated; no one has since taken the Discovery Institute seriously.

All three books, despite their regrettable titles, handle the basic story very well and recount some extraordinary moments. An OxyContin-addicted school-board member ranted on record: "Two thousand years ago someone died on a cross. Won't anyone take a stand for him?" and then denied that creationism had ever been discussed at board meetings. The school-board president claimed in his deposition that he did not know where the money came from to purchase the *Pandas* books, and then was shown the cheque from the other board member to his own father. Expert witness Barbara Forrest graphically showed that the authors of early drafts of *Pandas* had changed some 150 uses of terms such as 'creation' and 'creationist' to 'intelligent design' and 'design proponents', despite a 1987 Supreme Court decision ruling that 'creation science' was not science.

Where does the 'science' of intelligent design come from? Biochemist Michael Behe of Lehigh University in Bethlehem, Pennsylvania, is virtually the only scientist prominent in the movement; he has published popular books (for a review of the latest see *Nature* **445,** 1055–1056; 2007) but no demonstrable peer-reviewed research on intelligent design. Behe's notions of 'irreducible complexity' and the status of intelligent design as science were shredded by attorney Eric Rothschild, who got him to admit that under his own definition, astrology would qualify as science.

Conspicuously absent from the trial was William Dembski, the other pillar of intelligent-design 'research', who holds advanced degrees in maths and theology but none in science, and believes that intelligent design is the Logos of the Gospel of John restated in the language of information theory. His notion of 'specified complexity', a probabilistic filter that allegedly allows one to tell whether an event is so impossible that it requires supernatural explanation, has never demonstrably received peer review, although its description in his popular books (such as *No Free Lunch*, Rowman & Littlefield, 2001) has come in for withering criticism from actual mathematicians. Plaintiffs' attorneys were eager to take him apart, but Dembski exited the proceedings in a suspicious eleventh-hour dispute about having his own lawyer represent him in deposition.

All three books are entertaining and informative reads; on balance the nod goes to Humes for his comprehensive account, although Slack is concise and readable. Another book on the trial, by local reporter Lauri Lebo, is due out next year. It promises even more lively details of this perfect storm of religious intolerance, First Amendment violation and the never-ending assault on American science education. ■

Kevin Padian is professor of integrative biology and curator at the Museum of Paleontology, University of California, Berkeley. He is also president of the National Center for Science Education and was a *pro bono* expert witness in the Dover trial.

A lone voice in the greenhouse

The Callendar Effect: The Life and Work of Guy Stewart Callendar (1898–1964), the Scientist who Established the Carbon Dioxide Theory of Climate Change
by James Rodger Fleming
American Meteorological Society: 2007.
176 pp. $34.95

Robert J. Charlson

With so much written on the subject of carbon dioxide as a cause of climate change, it seems to have a settled history. But the word 'established' in this book's subtitle moved me to ask who actually came up with this now well-accepted theory, and what the basis is for James Rodger Fleming's claim that the subject of his biography holds this honour.

There seems to be little doubt that in 1827 Jean Baptiste Joseph Fourier first articulated the idea that "light finds less resistance in penetrating the air, than in repassing into the air when converted to non-luminous heat". In the 1860s, John Tyndall showed that CO_2 and water vapour both absorb and emit infrared radiation. Then, in 1896, Svante Arrhenius performed the first calculations of the sensitivity of Earth's temperature to changes in atmospheric CO_2. He went on to calculate (incorrectly) that it would take some 3,000 years for a 50% increase of its atmospheric content at the prevailing rate of coal consumption. He further calculated, on the basis of the measured infrared transmission of the atmosphere by Samuel Langley, that a 50% increase of CO_2 would warm Earth's surface by 3.4 °C.

So how did author Fleming come to state that the CO_2 theory was established by Callendar? It seems that this credit should be given to Fourier, Tyndall and Arrhenius.

Callendar's seminal paper, 'The Artificial Production of Carbon Dioxide and its Influence on Temperature', was published in 1938, nearly half a century after these nineteenth-century works. During the intervening period, serious doubts had developed about the importance of changing atmospheric CO_2 as a factor in Earth's climate and a cause of ice ages. Competing theories — changes in Earth's orbital geometry or in solar output, the role of the oceans, the attenuation of sunlight by volcanic dust, and spectroscopic considerations such as water vapour and CO_2 absorbing infrared light in the same spectral regions — had seemingly brought the CO_2-climate field into a 'deep eclipse'.

Guy Stewart Callendar revived the CO_2 theory of climate change.

UNIV. EAST ANGLIA ARCHIVE

Callendar's 1938 paper did not include a citation of Arrhenius's 1896 paper, although there are many parallels between the two. Callendar analysed just one set of data on atmospheric CO_2 content taken at Kew, near London, between 1898 and 1900. These data were taken near a source of CO_2 and were analytically very uncertain. From this analysis, he concluded that at around 1900 the free atmosphere over the North Atlantic region contained 274 ± 5 parts per million (p.p.m.) of CO_2. Then, after arguing that only a small fraction of the CO_2 from combustion of fossil fuels would dissolve in the ocean, he calculated from an estimated global production rate of CO_2 the amount that he thought would be there in 1936 (290 p.p.m.), 2000 (314–317), 2100 (346–358) and 2200 (373–396).

With a simple model of the absorption of infrared radiation, he worked out the amount of global warming to be expected from his predicted CO_2 levels, concluding that temperature would then have been increasing at a rate of about 0.03 °C per decade. Callendar's 1938 attribution of early twentieth-century warming to CO_2 increase might have been believable if global cooling had not ensued in the 1960s and 1970s.

His result was based on many assumptions and he used no contemporary CO_2 data on which to base his estimates. Nonetheless, his prediction was almost correct and, along with his 1958 paper — which included large amounts of CO_2 data (albeit of dubious quality) — his 1938 publication did rejuvenate the CO_2 theory of climate change. I doubt that this amounts to establishing the theory, but it came at a time when the fields of geochemistry and climate dynamics were ripe for stimulation, especially during the International Geophysical Year (1957–58). Shortly thereafter, Charles David Keeling presented accurate data, and the rest of the story is history.

Callendar's work on climate change is just part of the story Fleming tells about Callendar's life in this well written and especially well documented book. ■

Robert Charlson is in the Departments of Atmospheric Sciences and Chemistry, University of Washington, Seattle 98195, USA.

Ripples in relativity

Traveling at the Speed of Thought: Einstein and the Quest for Gravitational Waves
by Daniel Kennefick
Princeton University Press: 2007. 334 pp. $35, £22.95

Clifford Will

Gravitational physicists are eagerly awaiting the moment when gravitational-wave astronomy becomes a reality. More than half-a-billion US dollars have been sunk into ground-based, laser interferometric gravitational-wave observatories, and NASA and the European Space Agency are contemplating spending even more on the Laser Interferometer Space Antenna. Quite bold, when you consider that so far we have only indirect experimental evidence for gravitational waves. And, as Daniel Kennefick reminds us in his entertaining book, *Traveling at the Speed of Thought*, it was not so long ago that relativity theorists debated whether gravitational waves exist at all.

According to Einstein's general theory of relativity, the waves are ripples in the warpage or curvature of space-time that can be emitted by suitably moving masses; they travel with the same speed as light and, when they encounter a pair of masses, they cause them to move relative to each other, albeit by minuscule amounts.

Einstein in 1916 was the first to calculate the gravitational waves emitted by a system such as a rotating dumbbell. He showed that there are wave modes that can travel with arbitrary speed, which inspired Arthur Eddington in 1922 to coin the "speed of thought" phrase that forms the title of Kennefick's book.

We now understand these modes to be oscillations of the coordinates used to label the points in space-time, something that is allowed by the general covariance of Einstein's general theory of relativity but that has absolutely no physical consequences. Only waves of space-time curvature are physical. In hindsight, it seemed to take an inordinately long time — almost 40 years — to sort this out. Kennefick does a good job of describing the issues, the players and the many steps needed to get this and other issues squared away for good.

In 1936, Einstein thought he had proved the *non*-existence of gravitational waves. In what is now a legendary episode, he and his assistant Nathan Rosen at Princeton University, New Jersey, submitted the proof to the *Physical Review*, but Einstein was so appalled that the editor had dared to send it to an anonymous referee that he withdrew the paper and never published in that journal again.

But in Kennefick's telling, the story gets even more interesting. The referee's report, plus the results of a discussion between noted cosmologist H. P. Robertson and physicist Leopold Infeld, another of Einstein's assistants, may have convinced Einstein that the 'proof' was actually wrong. When he and Rosen later published their paper in the *Journal of the Franklin Institute*, it now claimed exactly the opposite — that the waves did exist. In a relativistic version of Watergate, people have wondered who was the 'deep throat' of *Physical Review*. Although suspicion naturally pointed towards Robertson, there was no clear proof. Kennefick helped find the smoking gun by discovering in Robertson's papers at the California Institute of Technology in Pasadena a carbon copy of the report, and by talking *Physical Review* into finding the old log books where then-editor John Tate had recorded sending the Einstein–Rosen paper to Robertson. For the experts, Kennefick reproduces Robertson's referee's report in an appendix.

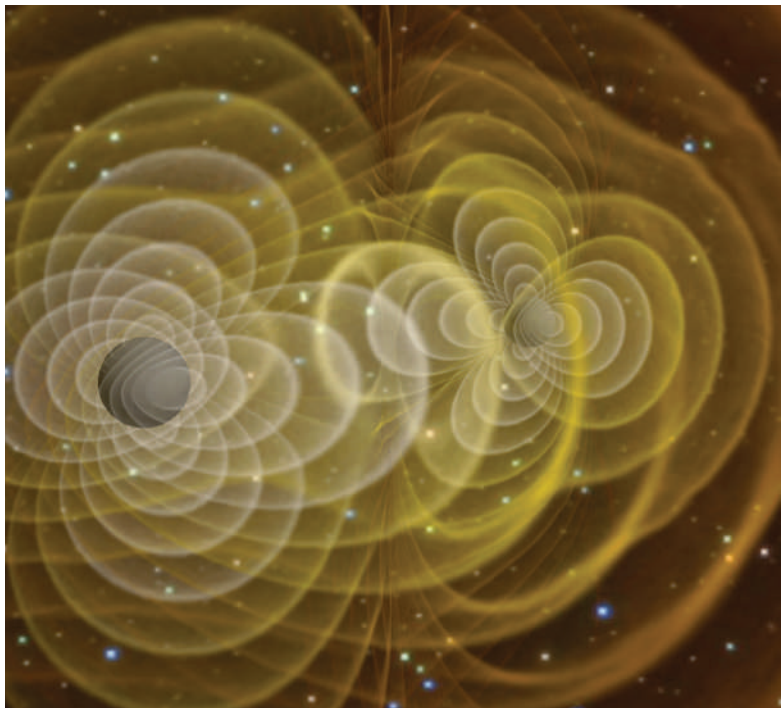
Kennefick also recounts another debate over whether gravitational waves would be emitted by binary star systems. Unlike a rotating dumbbell, the stars in a binary system are in free fall, and some relativists held that there would be no wave emission from such systems. Even among those who were convinced that there would be waves, there was debate about how to calculate their effects quantitatively. This is usually called the 'quadrupole formula controversy', named after the formula that gives the leading order effects. The debate over the validity of this formula raged rather strongly from the 1950s on, until the 1978 announcement by Joseph Taylor that the measured decay of the orbit of the 'binary pulsar' agreed with the quadrupole formula (today it agrees to 0.3%).

It is often said that history is written by the winners, and as a participant myself in the quadrupole controversies — on the winning side — I confess to having occasionally poohpoohed the role of the sceptics. The historian's role is to provide a richer perspective, and to elucidate the full dynamics of the story, taking advantage of hindsight, but also armed with interviews and a balanced reading of the documentary evidence. I must say that I learned much from Kennefick's retelling of the quadrupole wars and now appreciate better the viewpoints of the sceptics and their important contributions to the final resolution.

Kennefick, who is both a relativist and a science historian, writes in an engaging manner, although the book would be rather tough going for a lay reader. There is a fair amount of technical talk, plus some jargon from history and philosophy of science. Still, I recommend this book to readers with more than a passing interest in physics and its history. One thing I would have liked is a comparison of the controversies described in this book with other modern scientific controversies, and not just with controversies of the era of Galileo or Newton. Are the difficulties described here unique to general relativity, or are they typical whenever one is exploring uncharted scientific territory?

Today, theorists are calculating waves many orders of approximation beyond the simple quadrupole formula, to determine the precise signal emitted from the decaying orbits of black-hole binaries (one of the leading candidates for detection) possibly within a decade. We feel we are now on solid theoretical ground, so it is not a stretch to call this 'applied' general relativity. Kennefick's book reminds us of the complex path that brought us to this point. ■

Clifford Will is a professor of physics at Washington University, St Louis, Missouri 63130-4899, USA.



C. HENZE, NASA

Gravitational waves are ripples in the warpage or curvature of space-time.

SCIENCE IN CULTURE

B. MCMILLAN

Beijing bubbles

The Olympic Aquatic Centre will be housed in a giant block of foam.

Philip Ball

For architects in Beijing, anything is now possible. In the past two decades the city has been redesigned beyond recognition, and with the impending Olympic Games in 2008, no architectural innovation seems too grand or ambitious that China's energetic labour force cannot rise to the challenge. A swimming pool housed within a block of giant foam, then, is a breeze.

In the Water Cube — the Olympic Aquatic Centre designed by the Australian architectural company PTW in conjunction with the China State Construction and Engineering Company and design consultants Arup — an intricate network of steel struts forms a framework for transparent plastic panels that interlock into a mesh of bubble-like polyhedra (pictured), sliced into slabs 3.6 metres thick in the walls and 7.2 metres in the ceiling. There are 22,000 of these struts in the entire structure; laid end to end they would stretch for 90 kilometres.

The result, now largely built and scheduled for completion in October, may look flimsy, but the interlaced foam geometry offers surprising resilience: Arup claims that the low, flat-roofed enclosure could be stood on its end without deforming.

This lightweight, translucent shell provides heat insulation, acting as a kind of ornate greenhouse. According to the designers, a fifth of the incident solar energy will be trapped to heat the interior and the pools themselves. During the day, enough sunlight penetrates the roof to cut lighting costs by more than half, relative to conventional pool halls. And the aim is to capture and recycle 80% of the water falling on the roof or lost from the pools. The Water Cube thus embodies a spirit of water conservation that is becoming increasingly pertinent to northern China, where water scarcity is approaching a crisis that has prompted an awesomely ambitious and costly project to reroute water from the Yangtze River for more than 1,000 kilometres.

The bubbles of the roof and walls make up no ordinary foam. To the casual glance the network looks rather random and disorderly. But there is deep symmetry to it, for its 'unit cell' — the fundamental repeating element — consists of eight polyhedral cells. Six have 14 faces, the other two have 12, and these are comprised of regular hexagons and irregular pentagons with differing side lengths and angles. This might seem a strange choice — it is much more labour-intensive to create than one of the several networks that can be made from single, symmetrical cell shapes. The structure is, however, ostensibly that for which the total surface area of 'bubble' wall is the smallest known.

This structure was discovered in 1993 by physicists Denis Weaire and Robert Phelan at Trinity College in Dublin. Surprisingly, their complicated aggregate of cells fills space with slightly less total surface area than the unit cell long assumed to be optimal, a 14-sided polyhedron described by Lord Kelvin in 1887. This, in turn, trumped earlier candidates for an 'ideal' minimal foam cell, such as the rhombic dodecahedron and the truncated octahedron. There is still no formal mathematical proof, however, that Weaire and Phelan's solution cannot itself be bettered.

As triumphs of economy go, this one is rather Pyrrhic: the saving in surface area over Kelvin's foam is at best a mere 0.3%, which is scant recompense for the increased complexity of fabrication and assembly. In fact, the problem is made worse by the fact that the Water Cube design involves rotating the Weaire and Phelan foam and then taking a cut through it, creating 100 or so different 'part bubbles'.

Perhaps the real gain is aesthetic, the subtle blend of order and irregularity providing a tantalizing stimulus for the eye — which, as the designers point out, will be best enjoyed by competitors in the backstroke events. PTW architect Chris Bosse says that the designers liked the 'organic quality' of the Weaire and Phelan foam compared with the crystalline sterility of Kelvin's.

So although the polyhedral construction principle aligns the Water Cube with the faceted domes of Richard Buckminster Fuller, the notion of using an area-minimizing surface prompts comparison with the architecture of Frei Otto, whose tent-like canopies were designed by looking at the shapes of soap films draped over wire frames. It is fitting, then, that Otto's designs too were used for an Olympic swimming stadium, that in Munich in 1972.

The foam principle also makes the Water Cube part of the tradition of biomimetic architecture, evoking the bee's honeycomb (whose double layer of interlocking cells has also posed a long-standing problem in surface minimization) and the spittle bug's use of a foam to protect its larvae from predators. ■

Philip Ball is a consultant editor at *Nature*.

THEORETICAL PHYSICS

Walk the Planck

Where relativity and quantum mechanics clash, new laws of physics should emerge.

Giovanni Amelino-Camelia

The Planck scale is where general relativity and quantum mechanics should clash. It is the realm of the unfeasibly energetic and the unimaginably tiny. It is where the laws of nature are expected to achieve their highest level of elegance and simplicity — and where speculation abounds.

Albert Einstein's general theory of relativity follows a completely different logic from quantum mechanics. In relativity, observables evolve smoothly and deterministically. Quantum mechanics, in contrast, relies on quanta and probabilistic predictions.

Mostly, these differences are moot. Quantum mechanics describes the interactions of microscopic entities, such as the low-energy particles typically studied in laboratories. Here, general relativity can be ignored because the electroweak and electrostrong forces far outweigh gravitational forces. General relativity takes centre stage for the motions of macroscopic bodies such as planets. Here gravitational interactions dominate, because the bodies are composed of a large number of particles; the electroweak and electrostrong forces can be disregarded because they tend to average out.

Around the Planck scale — equivalent to 10^{19} gigaelectronvolts — things get tricky. For microscopic particles with energy this high or higher, both quantum-mechanical and general-relativistic effects come into play. Gravity enters the picture, in other words.

Right now we can only speculate about the outcome of such a clash — current particle accelerators produce energies 10^{16} times too low. Some believe that general relativity would adapt. At or near the Planck scale, they argue, gravitational interaction would become quantized. All interactions would be subject to a unified description governed by quantum mechanics. In this scenario, the Planck scale marks the beginning of a new regime for the laws of physics. It collapses our current theories' disorderly multiplicity of interactions into one law of interaction.

String theory provides a fascinating alternative scenario. It can accommodate a form of duality between the properties of a particle with energy a given amount below the Planck scale and the properties of a particle with energy a corresponding amount above the Planck scale. This may happen when the theory is formulated with extra space dimensions, undetected by our senses because of their tiny size. In this line of thinking, the Planck scale still marks the beginning of a new physics, but string theory's 'beyond-planckian' regime is just a mirror copy of the old 'sub-planckian' one.

There is a third, even more baffling, scenario. This posits that the Planck scale is the maximum possible energy of a fundamental particle, just as the speed of light is the maximum possible velocity. As a result, some new effects, such as certain particle-physics reactions forbidden in our current (pre-planckian) theories, would become more significant as the energies of the fundamental particles involved approach the Planck scale.

D. PARKINS

In string theory, the laws of physics above the Planck scale are a mirror image of those below it.

"The laws of physics at the Planck scale may remain nature's well-kept secret for a very long time."

Preliminary studies suggest that theories based on this Planck-scale energy limit might provide a 'fair' outcome for the clash. General relativity and quantum mechanics would at last be unified in such a way as to preserve some key features of both: the core concepts of relativity would be responsible for the emergence of the Planck-scale energy limit; quantum mechanics would have an important role in describing space-time structure.

Over the past few years, there has been an increased effort to devise ways to seek indirect experimental evidence to discriminate between these alternative scenarios. Some promising ideas are being developed, but it is a truly formidable challenge. The laws of physics at the Planck scale may remain nature's well-kept secret for a very long time. ■

Giovanni Amelino-Camelia is in the Department of Physics, University La Sapienza, and the Sezione Roma1 of the National Institute for Nuclear Physics, Piazzale Moro 2, Rome 00185, Italy.

CONCEPTS

NEWS & VIEWS

PALAEOGEOGRAPHY

Europe cut adrift

Philip Gibbard

The floor of the English Channel provides evidence for two catastrophic floods arising from the drainage of huge glacial lakes in the area of the southern North Sea. These megafloods made Britain what it is today.

"Floods in Channel, Continent Cut Off." This slight variation on a legendary headline in *The Times* — the original supposedly began with "Fog" — might exemplify Britain's notoriously insular view of the world. But it could also be an apt description of the events that led to Britain's becoming an island in the first place. The idea that a single, catastrophic flood was the root cause of this process is controversial. But, as they detail on page 342 of this issue, Gupta *et al.*[1] have just found the best evidence yet for not just one, but two such 'megafloods', in a bathymetric study of the morphology of the current Channel floor.

Standing on the southern English coast today looking at the Channel — or *La Manche*, from the French perspective — it is difficult to imagine that as little as 12,000 years ago the view would have been strikingly different. Then, instead of the sea, one would have been confronted by a vast shallow valley drained by a substantial river, larger than any in Europe today. This westward-flowing Channel river carried water not only from the rivers that currently enter the Channel, but also from those that now drain into the southern North Sea, among them the Rhine, Meuse, Thames and Scheldt[2] (Fig. 1).

This changed view was a product of global changes in climate. For most of its 50-million-year existence, the English Channel was a marine embayment. But throughout the past 2–3 million years, the build-up and decay of ice sheets on the continents have driven spectacular changes in global sea level. Today, for example, we live in an interglacial — a period characterized by relatively limited glaciation and therefore high sea levels. But at the peak of the last glaciation, 20,000 years ago, oceans across the globe stood about 100 metres below where they do now.

Throughout these climatic fluctuations, shallow areas such as the Channel basin and the North Sea have repeatedly emerged as dry land. Since its first emergence, about 1 million years ago, the Channel floor has been alternately modified by terrestrial and marine processes. During times of low sea level, a terrestrial drainage system began to form[3], feeding the Channel river aligned along the basin's axis.

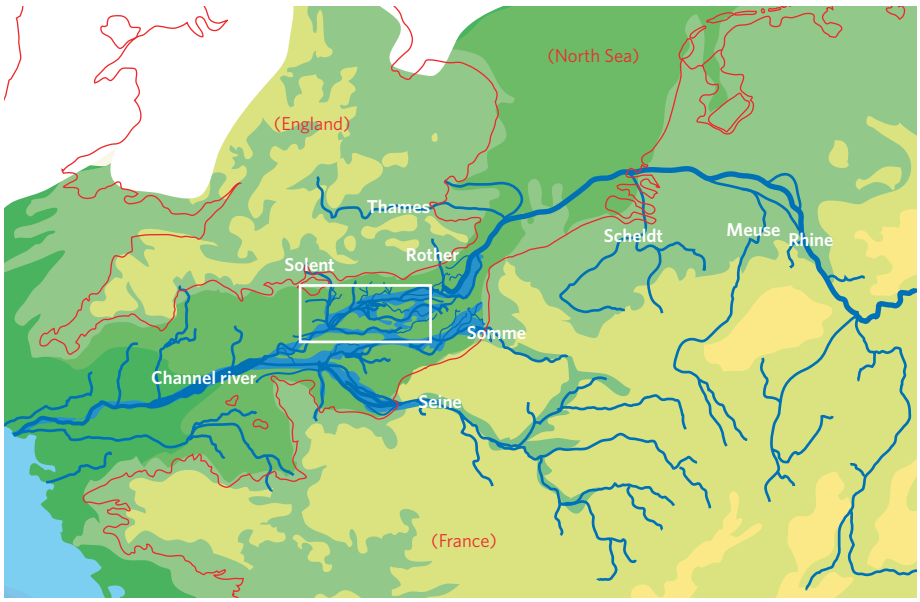


Figure 1 | A river runs through it. Where now the English Channel lies, flowed once a mighty river, draining not just the waterways of northern France (such as the Somme and the Seine) and the southern counties of England (the Solent and the Rother, for example), but also the Scheldt, the Rhine–Meuse and the Thames, which today discharge into the North Sea. Here, the Channel river system is shown during the last glacial maximum, around 20,000 years ago, as it was established after the second of two significant flood events proposed by Gupta *et al.*[1]. The area of the 'Northern Palaeovalley', whose morphology they studied, is marked by a square. (Map modified after ref. 9.)

Today, the smooth, shallow bedrock of the submerged Channel floor slopes gently for more than 500 km, from the Dover Strait in the east, its narrowest point, towards the Atlantic shelf margin that stretches between Brittany at the western extremity of France and Cornwall in England's extreme southwest. In the central and eastern Channel, this surface is dissected by a network of valleys, many of which are continuations of onshore river valleys: those of the Seine and Somme on the French side, for example, and of the palaeo-Solent and Rother on the English[3,4] (Fig. 1). These valleys join two substantial trunk valleys aligned along the axis of the Channel, the larger of the two being the 'Northern Palaeovalley'. It is from the morphology of this valley that Gupta and colleagues[1] draw their conclusions.

What they find there are distinctive features indicating that the valley formed in a catastrophic flooding event, rather than through normal fluvial erosion processes. The valley is unusually straight and wide, with prominent, streamlined margins and kilometre-scale grooving of the valley floor; the axis of the valley contains elongate islands characteristic of megaflood erosion; and the palaeo-Solent seems to form a 'hanging tributary' to the main valley, suggesting that the main valley's base level was suddenly lowered. What's more, the specific morphology of a bedrock bench at the valley margin, as well as evidence for an intervening period of normal fluvial erosion, indicates that during the evolution of the Channel at least two megafloods occurred, after 450,000 but sometime before 180,000 years ago.

Where did these floods come from? To answer this question, we must cast our eyes eastwards to the Dover Strait, or *Pas de Calais*. The origin of this narrow seaway linking the North Sea to the English Channel, cut through

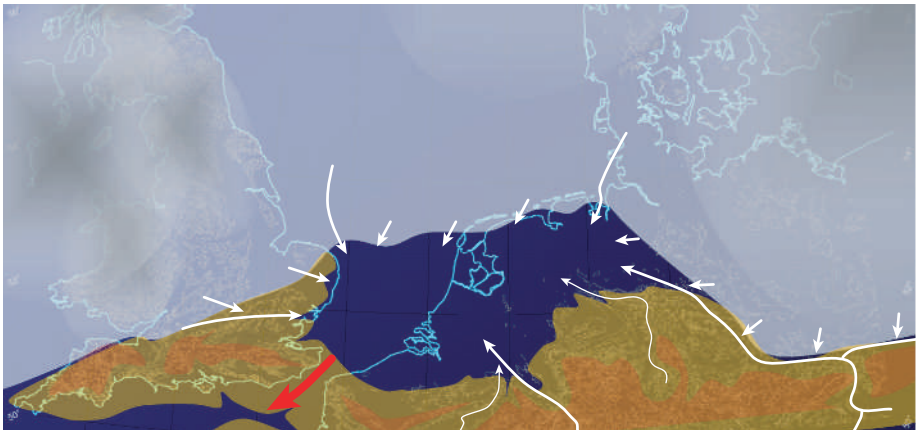


Figure 2 | Dammed lake. The first glacial lake in the area of the southern North Sea existed from around 450,000 to 400,000 years ago. This was dammed to the north by the extension of continental ice from Scandinavia to eastern England and central western Europe, and to the south by a band of higher land, the Weald–Artois anticlinal ridge, stretching from south-east England to northern France. The rivers entering the lake (white arrows) and the overspill at the Dover Strait into the Channel (red arrow) are shown. It was probably this overspill that initiated the first flood identified by Gupta *et al.*[1]. (Map modified, with permission, from ref. 6.)

a gently upfolding ridge of bedrock, has been a point of discussion for more than a century. The consensus now is that its formation was initiated some 450,000 years ago during the first major extension of a continental ice sheet into lowland central Europe and Britain[2,5]. The ice advanced across the emergent North Sea floor from southern Scandinavia, blocking rivers flowing northwards into the Atlantic Ocean and causing an immense glacial lake to develop in front of it, dammed by higher ground to the south and fed by the drainage of much of Western Europe (Fig. 2).

The lowest point of this dam, which was about 30 m above today's sea level[6], stood where the Strait of Dover now lies. Once this 30-km-wide bedrock barrier was overtopped some time around 425,000 years ago, the overflow quickly become torrential. The water would initially have followed existing stream valleys into the Northern Palaeovalley, but the deluge would quickly have overwhelmed these valleys, with the turbulent waters causing dramatic deepening and enlargement. This, in all probability, was the first flood identified by Gupta and colleagues[1].

It is no exaggeration to say that this first Channel flood was probably — on the basis of comparison of the landforms it sculpted with those formed by the Lake Missoula megaflood in the northwestern United States[4,7] — one of the largest ever identified. But although the Channel flood was only of comparable proportions to the Missoula flood, it had more profound long-term geographical consequences.

After the draining of the North Sea lake that resulted from this flood, the Thames and the Scheldt were realigned through the newly formed Dover Strait into the Channel river; the Rhine and Meuse, however, returned to the North Sea after the glaciers withdrew[5,6]. Their diversion to the south was delayed by another 200,000 years until, during a second period of significant glaciation, an ice sheet reached the central Netherlands and again dammed a lake in the southern North Sea[8]. This time, the level of the lake's water remained close to present sea levels, and its southern margin was not at the Dover Strait, but at a ridge further north. This was either an outcrop of bedrock, or possibly a barrier formed from moraine deposited during the previous glaciation.

The failure of this weaker barrier would have been immediate and catastrophic, and almost certainly released a vast volume of water that surged through the Dover Strait and thundered on into the Northern Palaeovalley. This event was probably the second and devastating flood so convincingly demonstrated by Gupta and colleagues[1]. By ensuring that the Dover Strait gap was greatly enlarged — almost to its present form — this single event finally sealed Britain's fate; during periods of high sea level, it would henceforth be an island. And during intervals of low sea level, such as the last glaciation, the Channel river would carry effectively half the drainage of western Europe to the Atlantic Ocean[2,3,9].

As Gupta *et al.* conclude[1], the implications of such significant palaeogeographical changes for plant and animal (including human) migration are manifold, culminating in the impoverishment of the British biota during the last and current interglacials, but also providing a land-bridge during glacial periods. In addition, the almost instantaneous release of huge volumes of freshwater into the Atlantic Ocean could have triggered changes in ocean circulation which might, in turn, have affected the climate of the whole North Atlantic region[10]. Britain's island story began here. ■

Philip Gibbard is in the Department of Geography, University of Cambridge, Downing Place, Cambridge CB2 3EN, UK.
e-mail: plg1@cam.ac.uk

1. Gupta, S., Collier, J. S., Palmer-Felgate, A. & Potter, G. *Nature* **448,** 342–345 (2007).
2. Gibbard, P. L. *Phil. Trans. R. Soc. Lond. B* **318,** 559–602 (1988).
3. Lericolais, G., Auffret, J.-P. & Bourillet, J.-F. *J. Quat. Sci.* **18,** 245–260 (2003).
4. Lautridou, J.-P. *et al. Bull. Soc. Géol. France* **170,** 545–558 (1999).
5. Gibbard, P. L. in *Island Britain: A Quaternary Perspective* (ed. Preece, R. C.) 15–26 (Geol. Soc., London, 1995).
6. Cohen, K. M., Gibbard, P. L. & Busschers, F. S. in *INQUA-SEQS Meeting Volume of Abstracts* (eds Dehnert, A. & Preusser, F.) 4 (INQUA-SEQS, Bern, 2005).
7. Smith, A. J. *Mar. Geol.* **64,** 65–75 (1985).
8. Busschers, F. S. *Unravelling the Rhine* Thesis, Vrije Univ., Amsterdam (2007).
9. Bourillet, J.-F., Reynaud, J.-Y., Baltzer, A. & Zaragosi, S. *J. Quat. Sci.* **18,** 261–282 (2003).
10. Mangerud, J., Astakhov, V., Jakobsson, M. & Svendsen, J. I. *J. Quat. Sci.* **16,** 773–777 (2001).

STEM CELLS

The magic brew

Janet Rossant

Researchers have engineered embryonic stem-like cells from normal mouse skin cells. If this method can be translated to humans, patient-specific stem cells could be made without the use of donated eggs or embryos.

Two reports in this issue[1,2] and one elsewhere[3] describe a seemingly simple method for changing differentiated adult cells into pluripotent stem cells. The 'gold-standard' test for pluripotency is the ability of a cell to contribute extensively to all adult cell types, including the germ line. The cells generated by these authors pass this test. The researchers introduced four gene-transcription factors into fibroblast cells originating from mouse skin, and specifically selected those cells that, in response to these factors, expressed genes indicative of a pluripotent state. Not only did all three teams manage to isolate cell lines that resembled mouse embryonic stem (ES) cells, but when they injected these cells into early embryos, the cells differentiated into all normal adult cell types.

A previous study[4] had shown that differentiated adult cells could be transformed into pluripotent cells when fused with ES cells. This hinted that factors found in ES cells might be essential to conferring pluripotency on other cells. However, the transcriptional profiles, modifications to chromatin (complexes of DNA and histone proteins) and DNA methylation status of ES cells are very different from

those of adult cells, indicating that pluripotency is probably under complex layers of control. It was, therefore, a surprise when Takahashi and Yamanaka[5] reported last year that they could produce cell lines with some of the properties of ES cells by introducing just four transcription factors associated with pluripotency — Oct3/4, Sox2, c-Myc and Klf4 — into mouse skin fibroblasts, and then selecting cells that expressed a marker of pluripotency, *Fbx15*, in response to these factors. These cells were called induced pluripotent stem (iPS) cells.

However, the generated iPS cells differed from ES cells in their gene-expression and DNA-methylation patterns. And when these cells were injected into normal mouse blastocysts (70–100-cell embryos), no live chimaeras — animals carrying cells throughout their bodies from both the original blastocyst and from iPS cells — were born.

Yamanaka and colleagues[1], as well as Wernig *et al.*[2] and Maherali *et al.*[3], surmised that if selection of iPS cells were based on the expression of genes that are more essential for pluripotency than *Fbx15*, this might improve the generation of truly pluripotent reprogrammed cells. They derived embryonic or adult fibroblasts that were engineered to express drug-selectable markers under the control of one or other of the two best-studied genes crucial for pluripotency — *Nanog* and *Pou5f1* (Fig. 1). After introducing the four factors (Oct3/4, Sox2, c-Myc and Klf4) by the technique of retroviral transfection, cells were subjected to drug selection. All three groups could derive stable cell lines. In terms of transcriptional, imprinting (expression of alleles predetermined by the parent from which they originated) and chromatin-modification profiles, these were essentially identical to ES cells. Maherali *et al.*[3] also report appropriate reactivation of the inactivated X chromosome in a female iPS cell line, and all authors present images of chimaeric mice, as well as evidence of germline transmission of the genetic content of iPS cells.

But questions remain about the exact sequence of molecular events that leads to this dramatic reprogramming, and whether additional changes, beyond the expression of the four transcription factors, are involved. The process of reprogramming is slow — colonies take up to 20 days to develop into real ES-like cells, and their frequency is quite low. Is this because only a few cells happen to express the right combination or levels of the four factors because of the random integrations of the retroviruses? Or are there additional events, perhaps associated with retroviral insertion, that are required for full transformation?

Continued expression of the four external factors may not be needed for the maintenance of iPS cells; in fact, these cells seem to show low-level expression of these factors. Maherali *et al.*[3] directly tested the requirement for continued expression of Oct3/4 and found that it was dispensable for iPS survival. Such tests need to be done for the other factors.

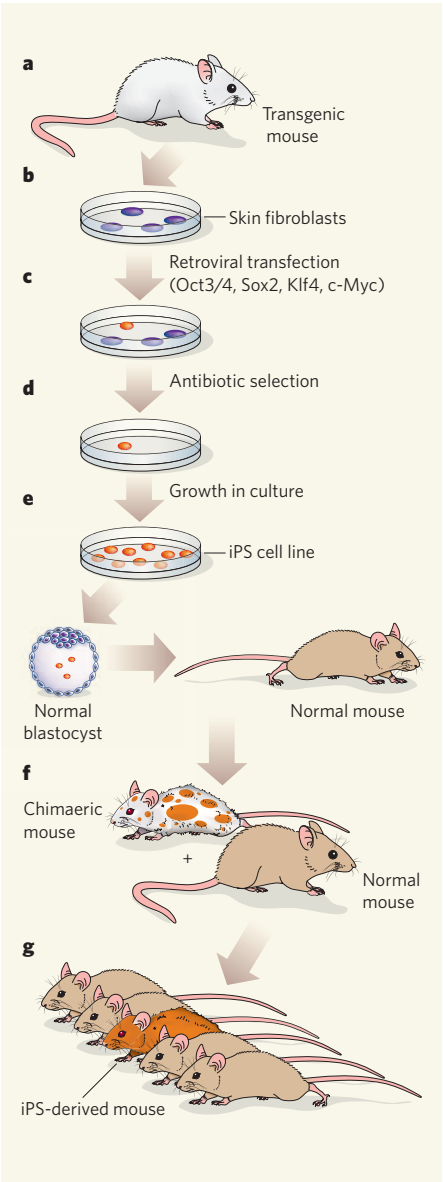


Figure 1 | Generating induced pluripotent stem cells. a, To turn adult skin fibroblasts into embryonic stem cells, researchers generated mice that carried a drug-selectable marker conferring resistance to either neomycin[2] or puromycin[1,3]. Resistance to these antibiotics was linked to expression of either *Pou5f1* (ref. 2) or *Nanog* (refs 1,3), which are markers of pluripotency. **b**, The authors then isolated fibroblasts from these genetically modified mice and, **c**, introduced genes for four transcription factors — Oct3/4, Sox2, Klf4 and c-Myc — into these cells by retroviral transfection. **d**, The transfected cells were subjected to appropriate drug selection. Cells that did not express the gene for drug resistance — that is, *Pou5f1* or *Nanog* — died. **e**, Thus, rare colonies that resembled embryonic stem cells were isolated, and expanded into stable induced pluripotent stem (iPS)-cell lines. **f**, When injected into blastocysts of normal mice, iPS cells could contribute to all cell types of the body, including the germ line. **g**, When the chimaeric animals resulting from these blastocysts were crossed with normal mice, this led either to the birth of live offspring carrying the genetic content of an iPS cell[1] or to development to the embryo stage[2,3].

Side effects of generating stem cells in this way are also problematic. Yamanaka *et al.*[1] found that 20% of the iPS-derived offspring developed tumours, presumably related to the activation of one of the transfected genes, *c-myc*. Some iPS-cell chimaeras have also developed tumours (R. Jaenisch, personal communication).

In addition, it is not known whether the same set of factors can reprogramme other, more specialized cell types. Clearly, the current methodology is more of a proof-of-principle than a fully rationalized series of molecular events that leads to reproducible reprogramming of adult cells. But it is an exciting proof-of-principle nonetheless. Multipotent progenitor cell lines — cells that can give rise to several, but not all, other cell types — have been developed from an increasing number of different fetal and adult sources; but none of these consistently shows the full pluripotency possessed by ES cells or these new iPS cells.

From proof-of-principle in mice to application in humans is still a leap. But since the publication of these studies online, media commentaries have focused on this possibility. The pluripotent nature of human ES cells holds enormous potential for future cell-based therapies for degenerative diseases and traumatic injuries. Use of human embryos to derive such cells remains controversial in many jurisdictions. The possibility of making cell lines with all of the properties of ES cells directly from non-controversial adult sources such as skin holds obvious appeal. It could also be a powerful way of making patient-specific stem cells to provide tissue-matched cells for therapy, and a source of cells for research into the pathogenesis of complex diseases.

Until now, the proposed method to make patient-specific ES-cell lines has been somatic-cell nuclear transfer (SCNT), or cloning. This involves reprogramming DNA from an adult cell by transplanting it into the cytoplasmic environment of an unfertilized egg[6] — or, more recently, of a newly fertilized egg[7]. ES cells derived through SCNT have been generated in mice, albeit at a fairly low frequency. In humans, the first reported success[8] from Woo Suk Hwang's group in South Korea is probably a parthenogenetic ES cell derived from a blastocyst, where the egg is activated without the sperm[9]; later claims by this group were found to be fraudulent. Descriptions of the generation of SCNT-derived primate ES cells at a recent stem-cell meeting in Cairns, Australia, indicate that it might be possible to derive human ES-cell lines using a similar method. However, it is still not clear whether human ES cells can be made at any reasonable frequency after SCNT. An efficient way of directly reprogramming adult cells, which also avoids the controversial use of donated eggs or embryos, would clearly be technically advantageous.

Will the same magic brew of molecular factors work to generate iPS cells in humans? Many

groups will certainly be rushing to test this. But translation from proof-of-principle to any therapy has many challenges. First, human cells would have to be given a built-in drug-selectable pluripotent marker by some efficient means. Second, potentially cancer-causing factors such as c-Myc must be avoided. Third, factors would need to be introduced by a method other than retroviral transfection; retroviruses can cause activation of cancer-causing genes and are therefore risky. Transient gene expression by direct introduction of membrane-permeable transcription factors into cells might be one way to achieve this, and screens for small molecules that can replace the gene products would also be useful. Despite these challenges, direct reprogramming of adult cells is clearly the way of the future, and promises to open up new frontiers in human biology and future therapy. ■

Janet Rossant is in the Program in Developmental and Stem Cell Biology, Hospital for Sick Children Research Institute, and the Department of Medical Genetics and Microbiology, University of Toronto, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada.
e-mail: janet.rossant@sickkids.ca

1. Okita, K., Ichisaka, T. & Yamanaka, S. *Nature* **448,** 313–317 (2007).
2. Wernig, M. *et al. Nature* **448,** 318–324 (2007).
3. Maherali, N. *et al. Cell Stem Cell* **1,** 55–70 (2007).
4. Tada, M., Takahama, Y., Abe, K., Nakatsuji, N. & Tada, T. *Curr. Biol.* **11,** 1553–1558 (2001).
5. Takahashi, K. & Yamanaka, S. *Cell* **126,** 663–676 (2006).
6. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. *Nature* **385,** 810–813 (1997).
7. Egli, D., Rosains, J., Birkhoff, G. & Eggan, K. *Nature* **447,** 679–685 (2007).
8. Hwang, W. S. *et al. Science* **303,** 1669–1674 (2004). Retraction Kennedy, D. *Science* **311,** 335 (2006).
9. Kitai, K. *et al. Cell Stem Cell* (in the press).

QUANTUM MECHANICS

Interference in the matter

Markus Kindermann

Like any particle, electrons are also waves that can interfere with each other. Remarkably, this interference can even happen between electrons from different sources that have never physically interacted.

That electrons are waves can be conclusively demonstrated by sending them through two parallel slits, and observing the interference patterns between them that result as they diffract. Such double-slit experiments led a recent ranking of the most beautiful experiments in the history of physics[1]. They were first performed with electrons in free space[2], but similar experiments performed later showed that the wave character also extends to electrons in small metallic conductors at very low temperatures[3]. Neder *et al.*[4], whose results appear on page 333 of this issue, have moved on to a next level: they have demonstrated compellingly that electron waves can interfere even if they originate from independent sources.

The statement that classical electromagnetic waves always add up ('superpose') no matter where they come from is quite intuitive, and typically taken for granted. The waves emitted by different light sources are thus able to interfere. This interference can even be observed for sources that are incoherent, or out of phase, through a phenomenon known as the Hanbury Brown–Twiss effect[5].

The same assertion made of electron waves is less obvious. The natural inclination is to describe each electron by its own matter wave that is independent of the waves associated with other electrons. And electrons with different origins are, one might think, evidently different.

This last assumption in fact turns out to be flawed. All electrons are intrinsically identical, and there is nothing besides their state at any particular moment that distinguishes two electrons (or any two elementary particles of the same kind) from each other. This is one of the fundamental postulates of quantum mechanics, and it follows that all electrons in the Universe are described by the same matter wave that invisibly connects them.

This strange fact has profound consequences. One of these is the phenomenon of 'anti-bunching': electrons tend to avoid getting too close to each other in space. Anti-bunching has been demonstrated with electrons in free space[6] and in electrical conductors[7,8]. But Neder and colleagues' painstaking demonstration of the interference of electron waves from independent sources[4] beautifully shows the fundamental mechanism underlying the anti-bunching phenomenon.

Quantum interference experiments in electrical conductors are typically performed[3] by splitting an electron beam from one source into two parts that propagate along different paths and merge again at a 'drain' (Fig. 1a). The phase difference between the two waves at the drain depends, through the so-called Aharonov–Bohm effect, on the magnetic flux between the two paths, which in turn is determined by the magnetic field in the area the two paths enclose. By tuning the magnetic flux, one can thus adjust the relative phases of the waves, and dictate whether the electron wave interferes constructively (if the peaks of the waves from each path arrive at the drain at around the same time) or destructively (if a peak and a trough arrive simultaneously). The electrical current at the drain accordingly oscillates as a function of the magnetic flux.

Neder *et al.*[4] have performed a similar experiment, but with two electron sources and drains (Fig. 1b). Their cleverly engineered interferometer, following a proposal by Samuelsson *et al.*[9], splits the source electron beams such that every emitted electron can reach both drain contacts, but the two parts of any one source wave cannot interfere with each other. Ordinary 'one-particle' interference is thus ruled out: at the drain contact, interference can take place only between waves originating from different sources.

As these sources are independent, the two waves should be emitted with random and uncorrelated phases. All effects of wave interference on the mean current in each of the

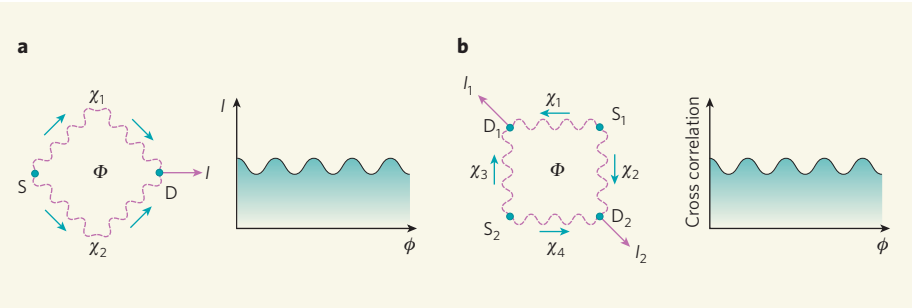


Figure 1 | One- and two-particle interferometers. a, In a single-electron interferometer, electrons enter from source S and choose between two paths to the drain contact D. The electron wave acquires different phases, χ_1 and χ_2, as it propagates along the two paths. In the presence of a magnetic flux, Φ, between the electron paths, the phase difference $\chi_1-\chi_2$ depends on Φ. The electrical current, I, into D oscillates as a function of Φ. **b**, In the two-electron interferometer demonstrated by Neder *et al.*[4], electrons enter from two sources, S_1 and S_2. Every electron is able to reach two drain contacts D_1 and D_2, through which mean currents $\langle I_1\rangle$ and $\langle I_2\rangle$ flow. A magnetic flux, Φ, between the electron paths again influences the phases χ_1, χ_2, χ_3 and χ_4 acquired by the propagating electron waves. Interference takes place only between waves originating from different source contacts; this causes oscillations with Φ of the cross-correlations between the currents in the two drains, defined as $\langle I_1I_2\rangle - \langle I_1\rangle\langle I_2\rangle$. (For each electron to be able to reach both drains, the quantum mechanics of electron scattering requires two additional, but unmonitored drain contacts; for simplicity these are not shown.)

drains should therefore average to zero. But the product of these two currents, as measured by a quantity known as a cross-correlation, has interference contributions where these random phases appear in pairs with opposite signs. This interference should not average out: it should oscillate as a function of the magnetic flux between all electron paths from the sources to the drains. This is precisely what Neder *et al.* observe.

Although the above discussion conveys the essence of the correct quantum-mechanical description, it is considerably simplified. Formally, the measured current cross-correlations are described by quantum-mechanical amplitudes of two-electron processes. Cross-correlations receive contributions from two processes: when an electron from source 1 travels to drain 1, while an electron from source 2 flows into drain 2; and when an electron from source 1 goes to drain 2, while that from source 2 finishes up in drain 1.

Because the electrons are indistinguishable, one cannot say which of these two things has happened, and the quantum-mechanical probability amplitudes for the two processes interfere. Because, however, the two events involve electron waves propagating along different paths, they do so with a phase difference that oscillates with the magnetic flux. This is thus two-particle interference that occurs without the two particles ever having physically interacted.

Experiments such as that of Neder and colleagues are pushing into a new and exciting area of nanoelectronics: the coherent control of many-particle quantum states. This control is one of the fundamental requisites for quantum information processing with electrons. Indeed, it has been proposed that one of the core resources for a quantum computer — entanglement, where the states of two remote particles become correlated at the quantum-mechanical level — can be created and detected in the next generation of such experiments[9,10]. Moreover, cross-correlation measurements reveal central aspects of the physics of interacting many-electron systems[11], including characteristics of their quantum state[12]. More beautiful fundamental physics, and more exciting applications, should be expected from this fascinating field of endeavour. ■

Markus Kindermann is in the School of Physics, Georgia Institute of Technology, 837 State Street, Atlanta, Georgia 30332-0430, USA.
e-mail: markus.kindermann@physics.gatech.edu

1. Crease, R. P. *Phys. World* **15(9),** 19–20 (2002).
2. Jönsson, C. *Z. Phys.* **161,** 454–474 (1961).
3. Webb, R. A., Washburn, S., Umbach, C. P. & Laibowitz, R. B. *Phys. Rev. Lett.* **54**, 2696–2699 (1985).
4. Neder, I. *et al. Nature* **448,** 333–337 (2007).
5. Hanbury Brown, R. & Twiss, R. Q. *Nature* **177,** 27–29 (1956).
6. Kiesel, H., Renz, A. & Hasselbach, F. *Nature* **418,** 392–394 (2002).
7. Henny, M. *et al. Science* **284,** 296–298 (1999).
8. Oliver, W. D., Kim, J., Liu, R. C. & Yamamoto, Y. *Science* **284,** 299–301 (1999).
9. Samuelsson, P., Sukhorukov, E. V. & Büttiker, M. *Phys. Rev. Lett.* **92,** 026805 (2004).
10. Beenakker, C. W. J., Emary, C., Kindermann, M. & van Velsen, J. L. *Phys. Rev. Lett.* **91,** 147901 (2003).
11. McClure, D. T. *et al. Phys. Rev. Lett.* **98,** 056801 (2007).
12. Kindermann, M. *Phys. Rev. Lett.* **96,** 240403 (2006).

NEUROBIOLOGY

New order for thought disorders

Lorna W. Role and David A. Talmage

Can we really learn about complex human psychiatric disorders through genetic manipulations in mice? Yes, according to studies of how altering the gene encoding neuregulin 1 affects signalling in the mouse brain.

Schizophrenia is a spectrum of disorders. Its causes lie in a complex interplay of genetic, prenatal and developmental factors, as well as precipitating events in later life. Given the complexity of this uniquely human disorder, surely it is hubris to think that poking around at a rodent gene or two could shed light on the processes underlying it? Results from Li *et al.*[1] and Woo *et al.*[2], published in *Neuron*, suggest not. Building on a link first made in 2002 between schizophrenia and a protein called neuregulin 1 (Nrg1) and its signalling partners, the ErbB receptors[3,4], these authors highlight how targeted manipulation of Nrg1 in mice can help to illuminate the workings of neural circuits gone awry.

The devastating array of psychotic, emotional and cognitive symptoms that comprise schizophrenia is thought to be caused by an imbalance in the fine tuning, or 'synaptic plasticity', of connections between neurons in the brain. This plasticity reflects the ability, in healthy individuals, to adjust, adapt and alter the levels of excitability of the myriad synapses and circuits that link different brain regions in a manner precisely coupled to ever-changing demands.

Li *et al.*[1] and Woo *et al.*[2] use a range of techniques to manipulate the levels of Nrg1–ErbB signalling with high precision in space and time. They examine the effects of this regulation on different electrical, neurochemical and morphological measures of synaptic plasticity and reach two fundamental conclusions: first, that synaptic plasticity requires precisely the correct level of Nrg1–ErbB signalling; and second, that the absolute levels of Nrg1–ErbB

50 YEARS AGO

When the history of the twentieth century is written, the year 1957 will surely be noted, *inter alia*, as the one when women really began to press their claims for equal career opportunities with men... In one sphere, however, they have made little progress. This, of course, is the world of industry... [Some] manufacturers will only take women applicants when there are no suitable male applicants... Women should think carefully before deciding on particular careers in industry. In such careers as general management, personnel management, and industrial medicine, a break for marriage and child-bearing should be no handicap and should enable the middle-aged woman to return to industry even more fitted for her job. In rapidly evolving specialist fields, however, where knowledge of chemistry, physics and other natural sciences are involved, she may find it easy to secure a post before marriage but difficult to return to it afterwards.

From *Nature* 20 July 1957.

100 YEARS AGO

Notwithstanding the much improved statistics recently issued by the Lunacy Commissioners, thoroughly satisfactory materials are still wanting for solving the question whether the prevalence of insanity is or is not increasing. The importance of the problem... imparts special interest to a paper by Mr. Noel A. Humphreys on the alleged increase of insanity... This paper shows in a striking manner the value of scientific statistics in checking crude figures. The author expresses a decided opinion that there is no absolute proof of actual increase of occurring insanity in England and Wales, and that the continued increase in the number and proportion of the registered and certified insane is due to changes in the degree and nature of mental unsoundness for which asylum treatment is considered necessary, and to the marked decline in the rate of discharge (including deaths) from asylums.

From *Nature* 18 July 1907.

50 & 100 YEARS AGO

NATURE|Vol 448|19 July 2007

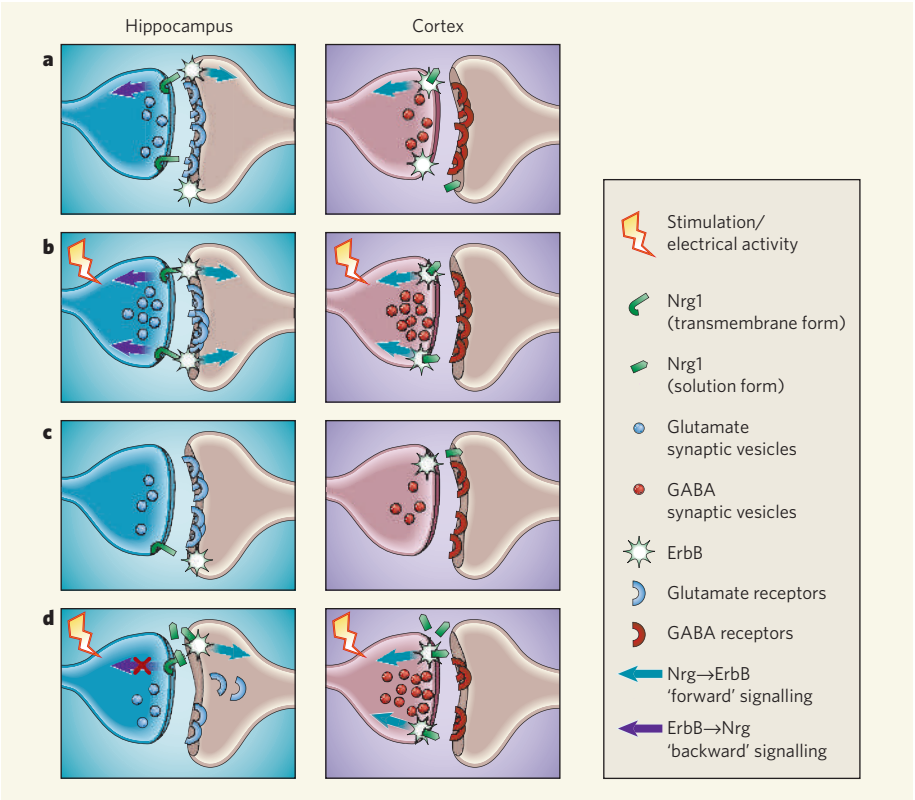


Figure 1 | Neuregulin's role in synaptic transmission. Together with other studies[5–8,11], the work of Li *et al.*[1] and Woo *et al.*[2] shows how Nrg1–ErbB signalling levels and particular patterns of neuronal activity regulate the strength (plasticity) of synaptic connections in the brain. In the hippocampus[1] (left panels), excitatory transmission is mediated by glutamate released from presynaptic neuron terminals and sensed by postsynaptic receptors; in the cortex[2] (right panels), inhibitory transmission is mediated by release of the neurotransmitter GABA. **a**, **b**, Long-term potentiation of the CA3–CA1 synaptic connection in the hippocampus requires electrical activity and both presynaptic Nrg1 and postsynaptic ErbB (ref. 1). Patterned stimulation (**b**) increases Nrg1–ErbB interactions and can affect both the level of glutamate released from the presynaptic terminal (L.W.R. and D.A.T., unpublished observations) and the magnitude of the response by the postsynaptic neuron[1,8]. A similar link between activity, Nrg1–ErbB signalling and GABA release exists in the cortex[2], although here the ErbB response is localized in the presynaptic terminal. **c**, If either Nrg1 or ErbB levels are reduced in the hippocampus or cortex, the combined action of activity and Nrg1–ErbB signalling is lost, and synaptic plasticity is not enhanced (refs 1, 2, 8; L.W.R. and D.A.T., unpublished observations). **d**, Increasing ErbB signalling by bathing neurons with soluble fragments of Nrg1, however, interferes with activity-dependent increases in synaptic strength, rather than further enhancing long-term potentiation[5,7,8,11]. In the hippocampus, this manipulation increases ErbB signalling in the postsynaptic neuron and would be predicted to decrease Nrg1–ErbB signalling in the presynaptic terminal, possibly generating an imbalance of signals across the synapse (ref. 6; L.W.R. and D.A.T., unpublished observations). It seems that either too little or too much signalling compromises synaptic response (Fig. 2).

signalling are regulated by different patterns and intensities of neural activity. Thus synaptic plasticity is held in critical balance by bidirectional Nrg1–ErbB signalling between neurons on either side of the synapse, and between these neurons and their supporting cells, called glia.

Li *et al.*[1] examine the role of Nrg1–ErbB signalling in the hippocampus, the brain region in which learning and memory is best established. The long-term activation ('potentiation') and depression of synaptic connections within the hippocampus — in particular between the CA1 and CA3 subregions studied by Li and colleagues — manifest themselves in changes in the number and strength of the synapses, and depend crucially on the activation of various types of receptor for their neurotransmitter, glutamate. The authors' selective manipulation of levels of Nrg1 expression in CA3 and ErbB in CA1 demonstrates that the long-term potentiation of connections between CA3 and CA1 requires both presynaptic Nrg1 and postsynaptic ErbB (Fig. 1). If either of these components is missing, glutamate-mediated transmission is impaired, and classic methods for inducing long-term potentiation elicit only transient activation.

Woo *et al.*[2] look at how Nrg1–ErbB signalling affects areas of the brain's cortex that are associated with working ('scratch-pad') memory and executive control of behaviour. Cortical activity is intricately controlled by both inhibitory and excitatory synaptic interactions, as well as by precise timing- and location-dependent modulation of synaptic inputs and neural activity. Woo and colleagues show that the extent of activity-dependent release of the neurotransmitter gamma-aminobutyric acid (GABA) at inhibitory synapses connecting neurons within the cortex to neurons projecting to other brain areas is determined by levels of Nrg1 signalling — this time mediated by presynaptic ErbB receptors (Fig. 1). Their study emphasizes that the degree to which the GABA-mediated neurons are electrically excited (depolarized) influences the effectiveness of Nrg1–ErbB interactions.

This last observation underscores one of the most intriguing aspects of the story: patterned neural activity is emerging as a regulator of neuregulin signalling, thus providing control of synapse maturation. How much, how fast, and in what pattern, waves of activity percolate through our nervous system essentially determines how well we learn, remember and move — in other words, all our essential behavioural characteristics. The early development, and subsequent plasticity, of circuits and behaviours depend on neural excitability, as well as other signalling molecules (neurotrophic factors). Is it possible that developmental disruptions of these activity patterns lead to altered levels of Nrg1–ErbB signalling, and therefore are a cause of complex thought disorders?

Despite the reductionist nature of these[1,2] and other (refs 5–11; L.W.R. and D.A.T., unpublished observations) studies, consideration of their findings, together with circuit analyses *in vivo*, encourages us to imagine how different levels and patterns of neural activity might influence the effectiveness of Nrg1–ErbB signalling interactions[9,10]. If Nrg1–ErbB signalling controls the efficacy and plasticity of other connections, and if neural activity typically regulates the extent of Nrg1–ErbB signalling, this could provide a link between neurotrophic factors and activity. That would close the control-system loop that governs experience-dependent changes in brain connectivity.

These stories are still far from complete, however, and there are some conflicts with aspects of previous reports[5,7,11]. These had found decreases in the number of glutamate receptors and inhibition of long-term potentiation following the addition of soluble fragments of neuregulin to *in vitro* preparations of hippocampal neurons (Fig. 1). The differences might be due to the experimental particulars — for example, the effect of adding soluble Nrg1 (refs 5, 7, 11) versus deleting endogenous Nrg1 or ErbB (refs 1, 2, 8; L.W.R. and D.A.T., unpublished observations) — but that explanation seems unsatisfying. An alternative model, the 'inverted-U model' (Fig. 2) might better resolve the inconsistencies. This model assumes that the initial conditions of a synaptic connection — including the local concentration of Nrg1- and ErbB-signalling components and the type(s) of neuregulin expressed, as well as the exact pattern of incoming neural activity — are all important determinants of subsequent shifts in synaptic strength.

The intricate mechanisms underlying the regulation of synaptic plasticity in the central nervous system remain to be resolved. Nevertheless,

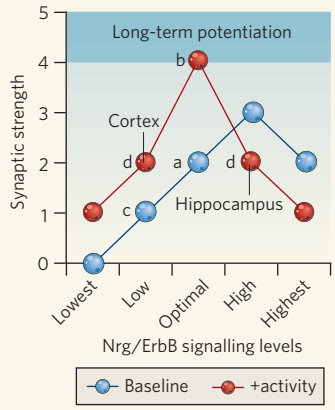


Figure 2 | An 'inverted-U model' of Nrg1 signalling. This model, based on all data to date (refs 1, 2, 5–11; L.W.R. and D.A.T., unpublished observations), attempts to resolve inconsistencies in the effects on long-term synaptic potentiation of manipulation of Nrg1 levels combined with different types of patterned stimulation. The model assumes that the initial conditions of a synaptic connection — including the local concentration of Nrg1 and ErbB signalling components, the type(s) of neuregulin expressed and the exact pattern of incoming neural activity — are all important determinants of subsequent shifts in synaptic strength from baseline activity to a long-term potentiated state. Receptor numbers either increase or decrease depending on whether the initial level of Nrg1–ErbB signalling is low, optimal or high. Lowering natural Nrg1 levels decreases potentiation, as does adding high concentrations of soluble Nrg1 to a synapse at normal signalling levels. Letters refer to panels in Fig. 1.

the recent studies[1,2] provide new perspectives and underscore how useful animal models can be, even for studying 'uniquely human' diseases of affect. They should encourage the use of further basic analyses to study the biological plausibility of genetic and environmental risk factors for susceptibility to psychiatric disorders, and thus to assess the therapeutic potential of treatment strategies. ■

Lorna W. Role is in the Department of Neurobiology and Behavior, and David A. Talmage is in the Department of Pharmacology, State University of New York at Stony Brook, Stony Brook, New York 11794, USA.
e-mail: Lorna.Role@Stonybrook.edu

1. Li, B., Woo, R.-S., Mei, L. & Malinow, R. *Neuron* **54,** 583–597 (2007).
2. Woo, R.-S. *et al. Neuron* **54,** 599–610 (2007).
3. Stefansson, H. *et al. Am. J. Hum. Genet.* **71,** 877–892 (2002).
4. Harrison, P. J. & Law, A. J. *Biol. Psychiatry* **60,** 132–140 (2006).
5. Huang, Y. Z. *et al. Neuron* **26,** 443–455 (2000).
6. Bao, J., Wolpowitz, D., Role, L. W. & Talmage, D. A. *J. Cell Biol.* **161,** 1133–1141 (2003).
7. Kwon, O. B., Longart, M., Vullhorst, D., Hoffman, D. A. & Buonanno, A. *J. Neurosci.* **25,** 9378–9383 (2005).
8. Bjarnadottir, M. *et al. J. Neurosci.* **27,** 4519–4529 (2007).
9. Buzsaki, G. *Neuron* **33,** 325–340 (2002).
10. Bartos, M., Vida, I. & Jonas, P. *Nature Rev. Neurosci.* **8,** 45–56 (2007).
11. Gu, Z., Jiang, Q., Fu, A. K., Ip, N. Y. & Yan, Z. *J. Neurosci.* **25,** 4974–4984 (2005).

PLANT BIOLOGY

Time for growth

Ghislain Breton and Steve A. Kay

Analyses of growth kinetics in seedlings reveal exquisite connections between the signalling pathways controlled by the circadian clock and by light, and illuminate the molecular mechanisms involved.

Using infrared light imaging to observe cell elongation in the dark, Nozue and colleagues (page 358 of this issue)[1] have made a fascinating discovery — that the growth rate of plant seedlings, specifically of a structure called the hypocotyl, is differentially regulated during a day–night cycle. Intriguingly, the maximal rates occur at dawn. To determine the molecular nature of this observation, the authors designed a series of experiments using several well-characterized mutants of *Arabidopsis thaliana*, a favoured subject in experimental plant biology. Their results allow them to separate the distinct contribution of light perception and the associated responses from that of the plant's circadian clock. Furthermore, they have identified two transcription factors — mediators of the production of messenger RNA from DNA — that regulate this cyclic mode of growth.

Genetic screens of *Arabidopsis* mutants have revealed the complex nature of growth regulation in the hypocotyl; this is a small region of about 20 epidermal cells in length that lies between the root and the embryonic leaves of young seedlings, and that grows mostly through longitudinal cell expansion[2]. Many mutations in genes involved in hormonal, light-perception and circadian pathways result in short or long hypocotyls[2,3]. Measurement of hypocotyl length in constant light or constant dark is commonly used to characterize light-signalling or clock mutants. Growth in constant darkness is thought to mimic the conditions experienced by seedlings that are emerging from the soil, and reaching for light at

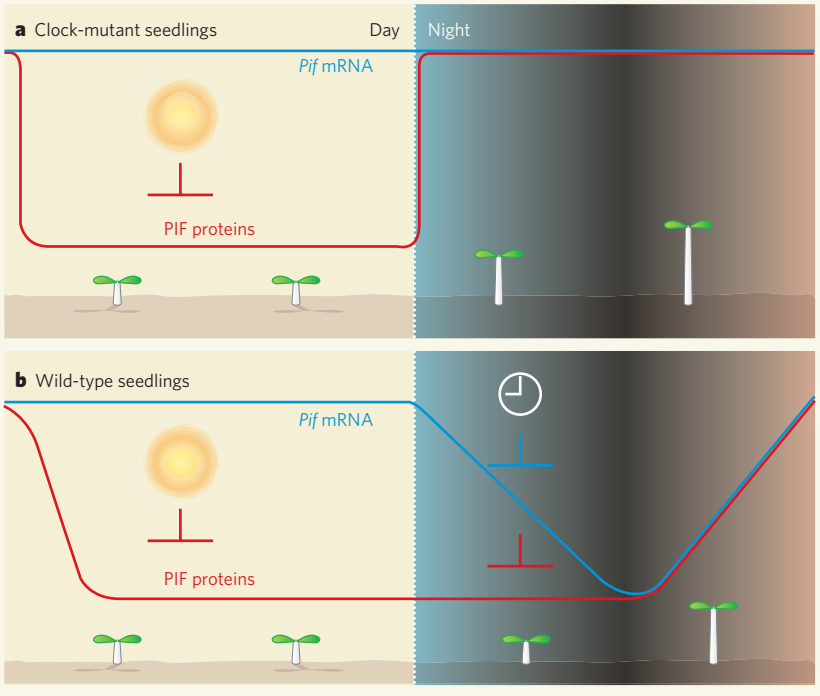


Figure 1 | From physiological observation to molecular mechanism. Nozue and colleagues' comparison[1] of hypocotyl growth rhythms in (**a**) clock mutants and (**b**) wild-type seedlings under light–dark cycles reveals the repressing effects of the circadian clock in the early night and of light during the day. **a**, Unregulated growth during the night; **b**, the normal growth pattern. Molecular studies identified two growth-promoting factors (PIF4 and PIF5), the messenger RNAs of which are normally clock-regulated (blue line) by a repressing action in the early night. In addition, the protein levels are reduced in a light-dependent manner (red line) possibly through a light-receptor-mediated mechanism. Thus, internal cues (clock) restrict transcriptional activation in the late night leading to hypocotyl elongation before dawn, whereas an external cue (sun) inhibits growth during the day by targeting the proteins for degradation. It is the coincidence of both cues that leads to the observed maximum growth at dawn.

the surface. In the dark, seedlings enter a form of development termed etiolation, in which most of the plants' resources are channelled into elongation of the hypocotyl. In contrast, seedlings grown in the light follow a different path, including the inhibition of etiolation and initiation of the greening process that enables light capture through photosynthesis[2,3].

Nozue and colleagues[1] have integrated study of these two conditions, which are normally considered separately, by measuring hypocotyl growth rate under diurnal conditions — that is, cycles of light and dark. They first noticed that, following a few days of non-consolidated growth, seedlings seem to tune their maximum growth at dawn. They showed that seedlings with specific defects in light perception had weak or no growth rhythms, suggesting that light signalling is essential for rhythmic growth under diurnal cycles.

To distinguish the role of the circadian clock on hypocotyl growth from that of light, they performed similar experiments using plants with clock defects. The output of the clock creates a temporal matrix that is used to drive overt rhythms, such as photosynthesis and protective mechanisms against cold at night, and can also serve to anticipate changes between day and night. One of the hallmarks of plant circadian clocks is their capacity to confer cycling behaviour under constant light conditions, and mutants with a disrupted clock have been used[4] to define a role for the clock in the timing of cell elongation.

Interestingly, Nozue *et al.* showed that, under conditions of diurnal cycles, maximal growth of mutants with an impaired clock occurred from dusk to dawn instead of being restricted to a few hours at dawn, suggesting that those seedlings were hyper-responsive to darkness. Additional experiments confirmed this observation, which implies that, under diurnal conditions, hypocotyl growth in wild-type (non-mutant) seedlings is partly controlled by light–dark transitions, whereas the circadian clock acts to suppress growth in the early part of the night.

What might be the molecular mechanism associated with growth control? To address this question, Nozue *et al.* carried out transcription profiling using whole-genome arrays. Taking advantage of the fact that clock mutants are hyper-responsive to darkness, they designed a strategy to find genes whose expression is associated with maximal growth rates. They identified two transcription factors — PIF4 and PIF5 (also known as PIL6), which belong to a family known as phytochrome-interacting factors (PIFs)[5] — to be good candidates as light-regulated components of the mechanism. Overexpression of each factor alone in *Arabidopsis* led to seedlings that were hyper-responsive to the dark. Furthermore, double mutants had a very short hypocotyl that didn't display any growth rhythmicity. These striking results confirm the growth-promoting nature of the PIF4 and PIF5 proteins.

Finally, two different experiments helped to determine how both genes might be regulated by light and the circadian clock. Expression-profiling experiments revealed that both are under circadian regulation and are expressed at high levels in seedlings without a functional clock (Fig. 1a). Because in wild-type seedlings the expression maxima are at dawn, and the minima are at the beginning of the night, the authors propose that, during the early night, growth suppression by the circadian clock must in part occur through transcriptional repression of the *PIF4* and *PIF5* genes (Fig. 1b). Further experiments with transgenic plants overexpressing PIF4 and PIF5 showed that the abundance of both proteins decreased in the light and increased in the dark. These results complete an elegant model of events, in which one of the essential processes involved in the reduction of growth rate after dawn is the light-dependent degradation of PIF4 and PIF5.

Questions remain, of course. Which factor negatively regulates *PIF4* and *PIF5* mRNA in the early night? Which molecular complex controls their degradation? And which genes are targeted by PIF4 and PIF5? More generally, there is the issue of whether Nozue *et al.*[1] have uncovered design principles that apply to growth regulation in other tissues, given that growth of the stem and leaf seem to be under similar control[6,7]. For the moment, however, publication of their discovery provides a considerable step forward in understanding the factors that shape the young seedling's quest for photons. ■

Ghislain Breton and Steve A. Kay are in the Department of Biochemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.
e-mails: gbreton@scripps.edu; stevek@scripps.edu

1. Nozue, K. *et al. Nature* **448,** 358–361 (2007).
2. Vandenbussche, F., Verbelen, J.-P. & Van Der Straeten, D. *Bioessays* **27,** 275–284 (2005).
3. Nozue, K. & Maloof, J. N. *Plant Cell Environ.* **29,** 396–408 (2006).
4. Dowson-Day, M. J. & Millar, A. J. *Plant J.* **17,** 63–71 (1999).
5. Khanna, R. *et al. Plant Cell* **16,** 3033–3044 (2004).
6. Wiese, A., Christ, M. M., Virnich, O., Schurr, U. & Walter, A. *New Phytol.* **174,** 752–761 (2007).
7. Jouve, L., Gaspar, T., Kevers, C., Greppin, H. & Degli Agosti, R. *Planta* **209,** 136–142 (1999).

CELL BIOLOGY

Caught in the traffic

Aparna Lakkaraju and Enrique Rodriguez-Boulan

In mice, deletion of the Rab8 protein disrupts organized molecular distribution to membranes of intestinal epithelial cells. Death by starvation follows, exactly as it does in humans with microvillus inclusion disease.

The gut, like other tubular structures in the body, is lined with a monolayer of epithelial cells, which forms the main interface between the external and internal environments of an organism. A defining characteristic of these cells is their polarity — that is, their apical and basolateral surfaces have different molecular compositions and different functions, allowing them to manage the different environments they face. This spatial asymmetry is achieved through sophisticated intracellular sorting mechanisms. Newly synthesized membrane proteins are packaged into specific carrier vesicles at the Golgi complex — the cell's 'post office' — and transported to their apical or basolateral destinations on the cell surface[1].

Crucial controllers of polarized membrane trafficking are a family of enzymes known as Rab GTPases, which are involved in vesicle targeting, docking and fusion. On page 366 of this issue, Sato *et al.*[2] report an unexpected function for one particular Rab GTPase, Rab8 — involvement in vesicle transport to the apical membrane of gut epithelial cells. These findings contribute to a better understanding of both the formation of the apical surface in epithelial cells and the pathogenesis of microvillus inclusion disease — a rare congenital human disorder for which no causative gene is known.

The most abundant type of gut epithelial cell, the enterocyte, has many finger-like projections at its apical surface called microvilli, which are rich in digestive enzymes known as hydrolases. These enzymes break down food into individual molecules (such as amino acids and sugars), which the enterocyte then transports into the blood.

Sato *et al.*[2] generated mice that lack the *Rab8* gene and found that these animals develop a lethal condition in which a type of vacuolar compartment rich in microvilli forms and apical hydrolases are mislocalized to these vesicles. The absence of hydrolases from their normal location at the apical surface resulted in death by starvation a few weeks after birth. This outcome, as well as pathological features of the condition, resembles that of human microvillus inclusion disease[3].

The work of Sato *et al.* is pioneering because it is the first study to examine Rab8 function in epithelial trafficking *in vivo*, and the team's observations raise some truly intriguing questions. First, why are only apical proteins affected by Rab8 deficiency, with the distribution of basolateral proteins being unchanged?

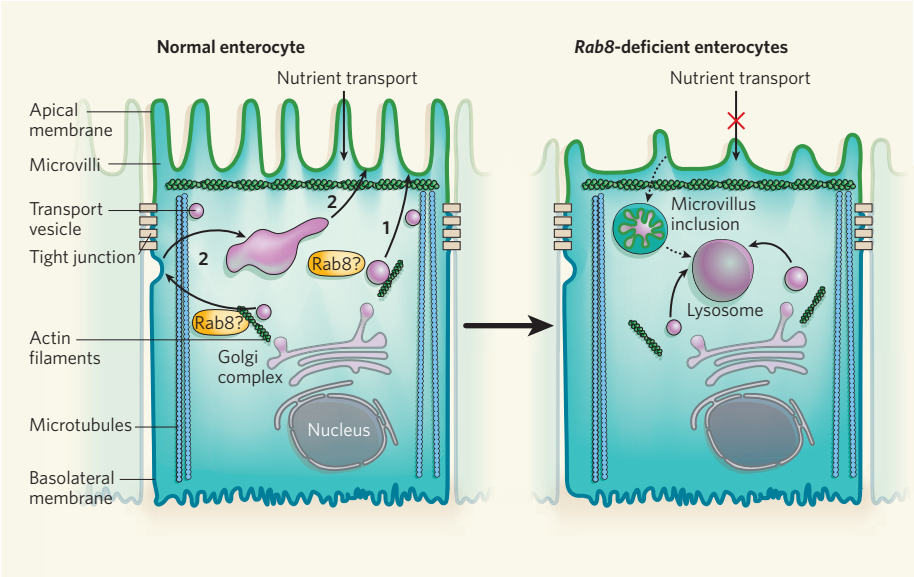


Figure 1 | The many functions of Rab8. Normally, epithelial cells have two distinct surfaces — apical and basolateral — which are separated by tight junctions. The apical surface of intestinal epithelial cells is rich in microvilli, which are responsible for nutrient absorption. Transport of apical membrane proteins such as intestinal hydrolases from the Golgi complex can follow a direct route (route 1) or an indirect route via the basolateral membrane (route 2). Sato *et al.*[2] suggest that transport via either or both of these routes might depend on the activity of the Rab8 GTPase. This protein regulates transport of vesicles by interacting with the actin- and microtubule-based cytoskeletal networks in the cell. The authors show that in intestinal cells that lack Rab8, apical membrane proteins are mislocalized to intracellular vacuoles containing microvilli, which may have been internalized from apical membranes by the process of endocytosis and which will be degraded in lysosomes by an unknown mechanism; consequently nutrient transport across the apical membrane is inhibited. These features recapitulate those found in microvillus inclusion disease, suggesting that RAB8 may function in the pathogenesis of this disorder.

This is surprising, because previously the only known function of Rab8 in epithelial trafficking was its involvement in basolateral transport. Earlier studies had shown that increased expression of mutant Rab8 disrupts the basolateral targeting of proteins in the kidney-derived MDCK cell line[4], and that Rab8 seems to function in trafficking pathways controlled by the basolateral adaptor protein AP1B (ref. 5).

A possible explanation for this discrepancy might be that intestinal cells use both a direct route (Golgi to cell membrane) and an indirect, transcytotic route (which entails a stopover at the basolateral membrane) to deliver apical proteins (Fig. 1). An outstanding question in epithelial cell biology is why certain apical proteins use the transcytotic route in some epithelial cells and a direct route in others[1].

It is thought that Rab8 regulates polarized trafficking by reorganizing microtubules and the actin cytoskeleton, which are involved, respectively, in long- and short-distance movement of transport vesicles within the cell. So a search for the molecular partners of Rab8 should provide essential information about the function of this protein in apical trafficking. Two such partners — optineurin[6] and FIP-2 (ref. 7) — have already been identified. These bind to the huntingtin protein[7], which regulates the function of the microtubule-associated motor protein dynein; dynein is essential for long-distance transport of some proteins to the apical membrane. Optineurin also binds to myosin VI (ref. 6), a Golgi-localized, actin-associated motor protein that participates in vesicular trafficking. It is possible that optineurin links Rab8 to myosin VI, thereby affecting short-distance movement along actin filaments and subsequent docking and/or fusion of Golgi-derived transport vesicles at the cell membrane; the Rab8–optineurin/FIP-2–huntingtin complex, meanwhile, might coordinate long-distance transport along microtubules.

Pertinent to the issue of epithelial polarity is the nature of the vacuolar compartment where intestinal apical hydrolases accumulate. Electron microscopy images of *rab8*-deficient small intestinal cells show that these enzymes are found in large vacuoles that often contain microvilli — a characteristic feature of microvillus inclusion disease[3].

Intracellular vacuoles containing microvilli have a long and interesting history. Remy[8] described these structures in cancerous epithelial cells as 'intracellular lumens' and speculated that cancer might disrupt a mechanism involved in the maintenance of the apical surface. Similar structures known as vacuolar apical compartments were observed[9] in epithelial cells cultured under conditions that prevented the formation of cell–cell contacts. On re-establishment of such contacts, vacuolar apical compartments fused with these areas of cell–cell contact, delivering microvilli, which later moved towards the apical surface of the cell. The formation of vacuolar apical compartments was thus suggested to be part of a normal mechanism for the generation of apical domains in developing epithelia. Indeed, intracellular vacuoles are present in the fetal intestine[10] and in endothelial cells during development[11], where they mediate lumen formation and morphogenesis.

What could be the link between the absence of Rab8, the mislocalization of apical hydrolases and the formation of intracellular vacuoles? Microtubules are a possibility. Several studies have shown microtubules to be crucial for the transport of apical proteins from the Golgi complex[12]. Both *in vivo* and *in vitro* experiments[12] have indicated that treating intestinal epithelial cells with microtubule-depolymerizing drugs results in the formation of intracellular microvilli-containing vacuoles that also serve as storage compartments for newly synthesized apical membrane proteins. These findings, together with the work of Sato *et al.*[2], raise the possibility that Rab8 acts as a 'receptor' to link Golgi-derived transport vesicles to the cytoskeleton, and that the absence of Rab8 not only misdirects these vesicles but also interferes with the biogenesis of the apical membrane.

Sato *et al.* tested three patients with microvillus inclusion disease. Although the authors couldn't identify any *RAB8* mutations in these patients, one patient had markedly low levels of both RAB8 protein and messenger RNA. The extremely low prevalence of this disease, together with its many characteristic features, makes unearthing its molecular basis a challenge.

Sato and colleagues[2] have opened a tantalizing treasure chest, the wonderful contents of which must be explored by future studies. In particular, it will be interesting to examine the organization of the actin cytoskeleton and the microtubule network in enterocytes lacking Rab8. These studies hold the key to insight into the acquisition and maintenance of epithelial polarity, cancer and intestinal disease. ■

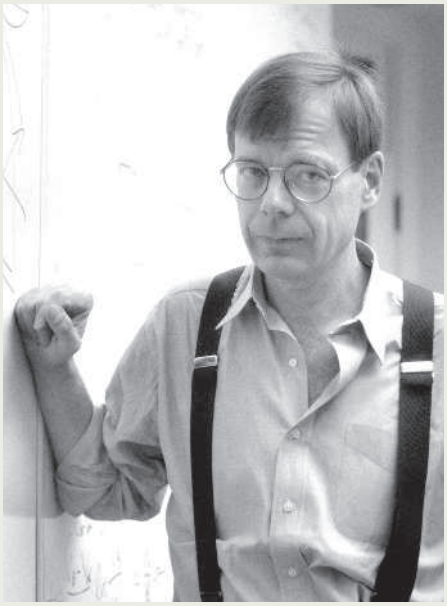
Aparna Lakkaraju and Enrique Rodriguez-Boulan are at the Margaret M. Dyson Vision Research Institute, Departments of Ophthalmology and Cell and Developmental Biology, Weill Medical College of Cornell University, New York, New York 10021, USA.
e-mail: boulan@med.cornell.edu

1. Rodriguez-Boulan, E., Kreitzer, G. & Müsch, A. *Nature Rev. Mol. Cell Biol.* **6,** 233–247 (2005).
2. Sato, T. *et al. Nature* **448,** 366–369 (2007).
3. Cutz, E. *et al. N. Engl. J. Med.* **320,** 646–651 (1989).
4. Huber, L. A. *et al. J. Cell Biol.* **123,** 35–45 (1993).
5. Ang, A. L., Fölsch, H., Koivisto, U.-M., Pypaert, M. & Mellman, I. *J. Cell Biol.* **163,** 339–350 (2003).
6. Sahlender, D. A. *et al. J. Cell Biol.* **169,** 285–295 (2005).
7. Hattula, K. & Peranen, J. *Curr. Biol.* **10,** 1603–1606 (2000).
8. Remy, L. *Biol. Cell* **56,** 97–105 (1986).
9. Vega-Salas, D. E., Salas, P. J. & Rodriguez-Boulan, E. *J. Cell Biol.* **107,** 1717–1728 (1988).
10. Colony, P. C. & Neutra, M. R. *Dev. Biol.* **97,** 349–363 (1983).
11. Kamei, M. *et al. Nature* **442,** 453–456 (2006).
12. Müsch, A. *Traffic* **5,** 1–9 (2004).

OBITUARY

F. Anthony Dahlen (1942–2007)

Pioneering and versatile theoretical geophysicist.



Tony Dahlen, probably the most important theoretical geophysicist of his generation, died on 3 June 2007, in Princeton, New Jersey. His seminal research on topics as far apart as seismology, Earth's rotation and the growth of mountains exerted a lasting influence on modern geophysics.

Dahlen was born in 1942 in American Falls, Idaho, while his father was serving in the US Navy in the Second World War, and moved with his reunited family to Winslow, Arizona, at the end of the war. There, in the shadow of the Barringer crater, he grew up searching for fossils and meteor fragments. Although it was a combined passion for geology, mathematics and physics, not American football, that earned him a Sloan Scholarship to the California Institute of Technology, he was nonetheless one of five members of the college football team who later became distinguished professors of geoscience.

His PhD research began in 1964, working with George Backus and Freeman Gilbert at the Scripps Institution of Oceanography of the University of California, San Diego. In the wake of the wartime boom in marine geophysical research, Scripps had become one of the principal US geoscience laboratories. It was a heady time, particularly for global seismology: two huge earthquakes, in Chile in 1960 and Alaska in 1964, had excited Earth's lowest 'eigenvibrations' — modes of oscillation in which the whole planet rings like a bell, with frequencies of a few cycles per hour. Two observed oscillations, with periods of 54 and 36 minutes, showed a splitting, akin to the Zeeman effect, in which an atom's energy levels split in a magnetic field. In this instance, it is the Coriolis force — an effect of Earth's rotation — together with Earth's slightly elliptical shape, that breaks symmetry and splits the spectral line into several closely separated oscillations.

Dahlen's thesis tracked the problem of the coupling of these modes. A fast worker, he could have graduated early in 1968, but decided to satisfy his broad interests by spending a further year sampling courses in other areas. In addition, in his first year at Scripps, he had caught the attention of a beautiful freshman one afternoon on the beach. She noticed he was reading a physics textbook, and asked if he would tutor her. Tony and Elisabeth Dahlen remained together until his death.

In 1970, Dahlen joined the faculty at Princeton University. His early work on eigenfrequencies provided a starting point for the idea that modes of oscillation split because of anomalous structures in Earth's interior, an idea that culminated in 1979 in a celebrated paper written with John Woodhouse. This work allowed seismologists to use eigenfrequencies for seismic 'tomography', to image small variations in Earth's elastic properties that are mostly caused by temperature variations. In the same period, Dahlen made important contributions to dislocation theory, which models the deformation due to earthquakes.This work led to a seminal paper on the energy balance of earthquakes.

Together with Martin Smith, Dahlen developed a linearized perturbation theory to describe the direction dependence of seismic-wave velocities in the interior of Earth. Today, this is the most practical means of visualizing the directions of the flow in Earth's mantle that is generated by convective processes. The two also determined the mantle's viscous response to stresses using the damping of a phenomenon known as the Chandler wobble, a 14-month precession of Earth's rotation. This followed on from earlier work in which Dahlen had developed a complete description of the effects of the oceans on variations in Earth's rotation.

In the 1980s, Dahlen moved on to work on the mechanics of the fold-and-thrust mountain belts and accretionary wedges that form at the margins of tectonic plates as they collide with each other. Collaborating with geologists Dan Davis and John Suppe, he showed how the formation of mountains in such regions can be explained in terms of a critical taper, mechanically analogous to the wedge of soil that forms in front of a bulldozer. Calculating the energy balance of such a system in western Taiwan, he showed that internal deformation contributes comparatively little energy to rock-transformation processes. Since then, submarine wedges such as those found off the Niger delta have been shown to act similarly. He also modelled the role of erosion of mountains and showed how this dominates the thermal evolution of mountain belts — a subject of much current interest.

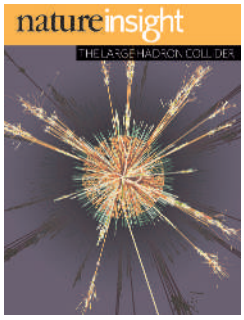
The influential textbook *Theoretical Global Seismology* (1998), written with his former student Jeroen Tromp, marked the culmination of three decades of research by Dahlen and others into low-frequency seismic waves. By the time the book came out, however, Dahlen had shifted his attention to the higher-frequency 'body waves'. These seismic waves reveal more detailed information about features deep within Earth if sampled sufficiently densely, although the theory underlying their propagation had barely evolved since the early twentieth century.

Dahlen rejected the paradigm that seismic waves propagate as narrow rays, as they do in classical optics, and he formulated an efficient computational strategy to take into account the diffraction of seismic waves. The advance allowed seismic-wave travel times to be interpreted more exactly as tomographic images. A first application by Raffaella Montelli led in 2003 to the serendipitous imaging of convecting plumes in the lower mantle — the first visual confirmation of a 30-year-old hypothesis that such plumes are the origin of ocean islands such as Tahiti and Hawaii.

Dahlen was an erudite and courteous scientist, unassuming and generous towards students and colleagues alike. Seismology is evolving into a science driven by huge quantities of data, with projects such as the USArray, funded by the National Science Foundation, providing hundreds of sensors to permit the sampling of a whole wavefield, rather than arrival times of individual rays. The scale of these endeavours demands the kind of theoretical advances that Dahlen provided, and the influence of his work is likely to be felt for years to come.

Although he received many honours, it is certainly the continuing relevance of his discoveries, and those of his students, that would have pleased Tony Dahlen most. Throughout his life he was driven by his love for science; his greatest pleasure was that his son Alex also decided to pursue a career in science.

Guust Nolet
Guust Nolet is in the Department of Geosciences, 320 Guyot Hall, Princeton University, Princeton, New Jersey 08544, USA.
e-mail: nolet@princeton.edu



Cover illustration
Simulated outcome of a proton–proton collision at the LHC, showing the tracks left by particles. (Courtesy of CERN)

THE LARGE HADRON COLLIDER

We are on the threshold of a new era in particle-physics research. In 2008, the Large Hadron Collider (LHC) — the highest-energy accelerator ever built — will come into operation at CERN, the European laboratory that straddles the French–Swiss border near Geneva.

In the debris of the collisions between protons in the LHC, physicists are hoping for a sign. Hypotheses such as the Higgs mechanism (and associated particle) and supersymmetry remain to be proved. The high energy of the LHC collisions will give the best ever 'reach': that is, the best chance yet of finding the decisive signature of a Higgs or supersymmetric particle.

In looking forward, we should not forget what lies behind us — the phenomenal success of the standard model of particle physics over the past three decades. The standard model has been tested to a greater degree of precision than any other model in science and has withstood every challenge. But it is incomplete, and the search for the missing pieces of the puzzle is the prime motivation for building the LHC.

The LHC programme is, however, much wider than a search for the Higgs. Alongside the 'general-purpose' detectors, known as ATLAS and CMS, the LHCb experiment will analyse the production of bottom quarks in LHC collisions. This rich system is the key to a better understanding of the phenomenon of CP violation and its connection to the dominance of matter over antimatter in the Universe. In addition, during dedicated runs in which lead ions will collide in the LHC instead of protons, the ALICE experiment will study a phase of matter called quark–gluon plasma, which might have existed shortly after the Big Bang.

This is the story told in this Insight: how the standard model was developed and tested; how it was agreed to build the LHC; how the programme has been realized through decades of effort by thousands of scientists — and how marvellous the rewards might be.

Alison Wright, Chief Editor, *Nature Physics*
Richard Webb, Senior Editor, *Nature* News & Views

nature insight

NATURE|Vol 448|19 July 2007|doi:10.1038/nature06073

The standard model of particle physics

The model describes the interplay of three forces — electromagnetic, weak and strong — and 12 elementary matter particles. (Gravity is not included.) Each force is mediated by the exchange of carrier particles: the photon, *W* or *Z* boson, or the gluon, as shown. Matter particles are divided into leptons and quarks, and, according to their mass hierarchy, line up into three 'generations'. Matter particles also have antimatter equivalents — such as the positron, which is an antielectron. For quarks, the antiparticles are typically represented by a bar placed over the letter that symbolizes them (for example, $\bar{u}$ is the antiparticle of the *u* quark). Collections of quarks and antiquarks form other, composite particles known as hadrons, a selection of which are shown. Hadrons are divided into mesons and baryons: mesons comprise a quark and an antiquark; baryons (including the proton and the neutron) are three-quark states.

Force	Carrier
Electromagnetic	Photon
Weak	$W^{\pm}$, Z
Strong	Gluon

	Leptons		Quarks	
	Electric charge −1	Electric charge 0	Electric charge $+^2/_3$	Electric charge $-^1/_3$
Three 'generations'	Electron e^-	Electron neutrino ν_e	Up u	Down d
	Muon μ^-	Muon neutrino ν_μ	Charm c	Strange s
	Tau τ^-	Tau neutrino ν_τ	Top t	Bottom b

Increasing mass

Hadrons			
Mesons Quark–antiquark states		**Baryons** Three-quark states	
π / ρ	$u\bar{d}$ / $\bar{u}d$ / $(u\bar{u} - d\bar{d})$	Proton (p), neutron (n)	uud, udd
η / φ	$(u\bar{u} - d\bar{d} + s\bar{s})$	Δ^{++}, Δ^-	uuu, ddd
K^0, $\bar{K}^0$	$d\bar{s}$, $\bar{d}s$	Λ^0	uds
K^+, K^-	$u\bar{s}$, $\bar{u}s$	Λ_c	udc
D^0, $\bar{D}^0$	$c\bar{u}$, $\bar{c}u$	Σ^0	uds
D^+, D^-	$c\bar{d}$, $\bar{c}d$	Σ^+, Σ^-	uus, dds
B^0, B_s^0	$d\bar{b}$, $s\bar{b}$	Ξ^0	uss
B^+, B^-	$u\bar{b}$, $\bar{u}b$	Ξ^-	dss
J/ψ	$c\bar{c}$	Ω^-	sss

The making of the standard model

Gerard 't Hooft

A seemingly temporary solution to almost a century of questions has become one of physics' greatest successes.

The standard model of particle physics is more than a model. It is a detailed theory that encompasses nearly all that is known about the subatomic particles and forces in a concise set of principles and equations. The extensive research that culminated in this model includes numerous small and large triumphs. Extremely delicate experiments, as well as tedious theoretical calculations — demanding the utmost of human ingenuity — have been essential to achieve this success.

Prehistory

The beginning of the twentieth century was marked by the advent of two new theories in physics[1]. First, Albert Einstein had the remarkable insight that the laws of mechanics can be adjusted to reflect the principle of relativity of motion, despite the fact that light is transmitted at a finite speed. His theoretical construction was called the special theory of relativity, and for the first time, it was evident that purely theoretical, logical arguments can revolutionize our view of nature. The second theory originated from attempts to subject the laws found by James Maxwell for the continuum of electric and magnetic fields to the laws of statistical mechanics. It was Max Planck who first understood how to solve this: the only way to understand how heat can generate radiation is to assume that energy must be quantized. This theory became known as quantum mechanics.

At first, it was thought that quantum mechanics would apply only to atoms and the radiation emitted by their electrons. But, gradually, it became clear that the laws of quantum mechanics had to be completely universal to make sense. This idea of universality was in common with Einstein's theories of relativity. In particular, quantum mechanics had to apply not only to electrons but also to the particles that reside in atomic nuclei.

It was clear, right from the beginning, that the two new theoretical constructions would need to be combined into one. The vast amounts of energy found to inhabit atomic nuclei implied that 'relativistic quantum mechanics' had to apply to atomic nuclei in particular. Thus, a new problem became evident and soon garnered worldwide attention: how is quantum mechanics reconciled with special relativity? This question kept physicists busy for most of the rest of the century, and it was not completely answered until the standard model saw the light of day.

The early days

By 1969, the reconciliation of quantum mechanics with special relativity was still a central issue[2], but much more had been discovered through experimental observation[3]. Matter particles (see page 270) had been divided into leptons and hadrons. The known leptons were the electron, the muon and their two neutrinos (these last assumed to be massless); hadrons, such as protons and pions, obeyed the conservation laws of quantum numbers known as 'strangeness' and 'isospin'. Hadrons are divided into mesons, which can be described loosely as an association of a quark and an antiquark, and baryons, which can be simply depicted as being made up of either three quarks or three antiquarks. The symmetry of strong interactions between subatomic particles was known to be approximated by the 'eightfold way' (Fig. 1). And it seemed that all hadrons had infinite series of excited states, in which angular momentum was bounded by the square of the mass measured in units of ~1 gigaelectronvolt (Fig. 2). This feature of all hadrons was telling us something important about strong interactions, but the first attempts to understand it consisted of rather abstract formalisms.

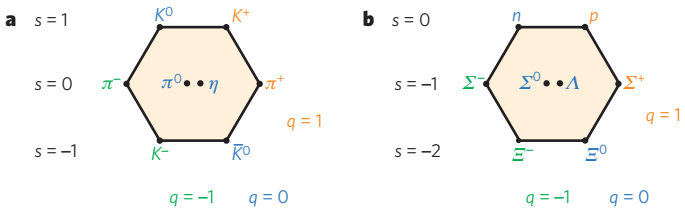


Figure 1 | The eightfold way. Spin-zero mesons (**a**) and spin-half baryons (**b**) can be grouped according to their electric charge, *q*, and strangeness, *s*, to form octets (which are now understood to represent the flavour symmetries between the quark constituents of both mesons and baryons).

It was also known that there are weak forces and electromagnetic forces, to which subatomic particles owe some of their properties. However, only the electromagnetic force was understood in sufficient detail for extremely precise calculations to be checked against accurate experimental observations. Theorists had tried to devise methods to subject not only the electromagnetic force but also other forces to the laws of quantum mechanics and special relativity. Despite their efforts over nearly 50 years, attempts to improve this 'quantum field theory' to include weak interactions failed bitterly. And describing the strong interactions between mesons and baryons drove them to despair.

The theorists at that time therefore concluded that quantum field theory should be dismissed as a possible way of addressing the dynamics of particle interactions. We now know that this was a misjudgement. Their mistrust of quantum fields was, however, understandable: in all known quantum field systems, there were divergences in the high-energy domain, making these systems unsuitable for describing strong interactions. Yet it was clear that strong interactions, such as those that hold a nucleus together, do exist. The error made by the theorists was that this 'bad' high-energy behaviour was thought to be an unavoidable, universal feature of all quantum field theories[4].

Because of this widespread objection to quantum field theories, few theorists ventured to investigate field theoretical methods. They should have realized that their objections could be swept away when the forces are weak. Indeed, the weak force was the first subatomic force to be formulated using the new 'gauge theories'[2]. Such theories had been proposed in 1954 by Chen Ning Yang and Robert Mills (Fig. 3), who were inspired by the fact that the two basic forces of nature that were well understood, gravity and electromagnetism, were both based on the principle of local gauge invariance: that is, that symmetry transformations can be performed in one region of space-time without affecting what happens in another. This beautiful idea got off to a slow start, even after Peter Higgs, François Englert and Robert Brout realized in 1964 how the structure of the vacuum can be modified by the field of a scalar (spin-zero) particle, which came to be called the Higgs particle. With the inclusion of the Higgs particle, the Yang–Mills field equations could

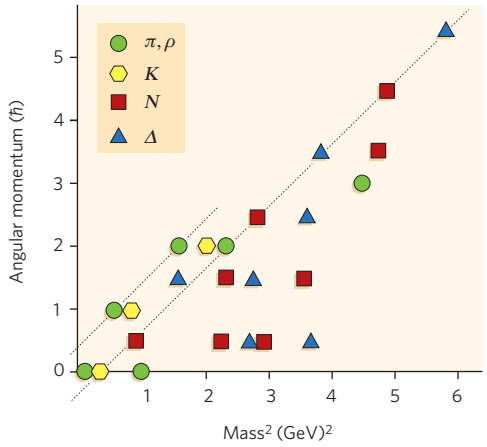


Figure 2 | A hint at the nature of the strong force. All strongly interacting particles and their excited states seem to have an angular momentum (in units of the reduced Planck constant $\hbar$) that is less or about equal to the square of their mass (measured in gigaelectronvolts, GeV). The limits for various particle species form lines that seem to be straight and parallel. *N*, nucleon (which includes neutrons and protons).

now be used to describe the weak force accurately; the force would be carried by the quanta of the Yang–Mills field, which had gained mass by this 'Brout–Englert–Higgs mechanism'. Reasonably realistic models in which exactly this happens were proposed by Abdus Salam, Sheldon Glashow and Steven Weinberg in the 1960s.

The 1970s

In 1971, Martinus Veltman and I demonstrated that it is exactly these theories (in which the mass of the Yang–Mills gauge quanta is attributed to the field of a Higgs particle) that are 'renormalizable', and it seems that this was all that was needed for a full rehabilitation of quantum field theory to begin[4]. Renormalization is the mathematical description of the requirement for distinguishing, at a fundamental level, the algebraic mass terms and coupling terms in the equations from the actual physical masses and charges of the particles. The choice of values for these algebraic parameters depends crucially on the smallest distance scales taken into account in the theory. So, if it were insisted that all particles are truly point-like — that is, the smallest distance scale should be zero — then these algebraic parameters would need to be infinite. The infinite interactions were needed to cancel the infinitely strong self-interactions of particles that the equations inevitably lead to. But the mathematical procedure of cancelling infinite forces against one another needed to be performed with considerable care. Many theorists did not understand what was going on and aired their strong suspicions that 'all this' had to be 'rubbish'.

We were learning not only how to construct realistic and logically coherent models but also how to study the behaviour of these theories over short distances by using the concept of the 'renormalization group'. Introduced by Ernst Stückelberg and André Petermann in 1953, this mathematical procedure allows one to go from one distance scale to another. It is used both in condensed-matter theory and in elementary particle physics, for which it was pioneered by Curtis Callan and Kurt Symanzik. A function that can be computed for every theory, named β-function by Callan and Symanzik, determines what might happen: if β is positive, the strengths of the couplings are increased at shorter distances; if β is negative, they are weakened. The error that I mentioned earlier was that all quantum field theories were thought to have positive β-functions. Indeed, it was claimed that this could be proved. Owing to various miscommunications, earlier calculations that yielded negative β-functions (including calculations by me) were systematically ignored, until in 1973, David Politzer, David Gross and Frank Wilczek published their findings that, for Yang–Mills theories, β is generally negative. Therefore, the strength of interactions would be reduced at short distances, making them controllable — a property that was named asymptotic freedom. Until this point, Yang–Mills theories had been understood to describe only electromagnetic and weak interactions. But the discovery of asymptotic freedom immediately turned Yang–Mills theory into a prime candidate for describing strong interactions as well.

In fact, experimental observations had been pointing in the same direction. A Yang–Mills structure not only fitted beautifully with the algebraic symmetries that had been established for the strong force (such as the eightfold way) but also could be deduced from observations made at the Stanford Linear Accelerator Center (SLAC), in California, where strong interactions seemed to show scaling behaviour, as though their strength diminished at short distances (known as Bjorken scaling)[4]. Indeed, theorists had concluded that no quantum field theory would be suitable for the strong force — until the asymptotic freedom of Yang–Mills fields was uncovered.

Basically, Yang–Mills fields are a generalization of the electromagnetic field, for which Maxwell had determined the equations a century earlier. Particles carry a generalized type of electric charge, allowing them not only to be accelerated by the Yang–Mills fields but also to be transmuted into other kinds of particle under the influence of these fields. Thus, electrons can transform into neutrinos, protons into neutrons and so on, as a result of the weak force. The strong force is understood as a new kind of field acting on quarks, which are the building blocks of protons and neutrons inside the atomic nuclei. In addition to ordinary electric charges, quarks also carry a threefold charge, which is reminiscent of colour vision (and hence they are usually called red, green and blue). For this reason, the Yang–Mills theory for the strong force is called quantum chromodynamics, the Greek word *chromos* meaning colour.

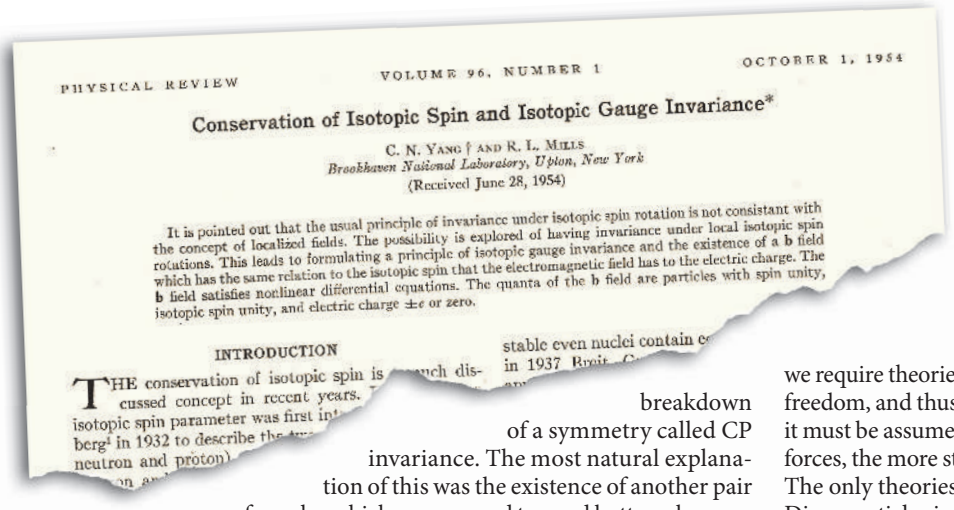
Getting the details right

For the first time, all known particles and all known forces between them could be cast in a single model. This model described three closely related Yang–Mills systems for the three major forces (strong, weak and electromagnetic), one Higgs field and several matter fields. These matter fields were Dirac fields, describing the four known leptons and the three known quarks (up, down and strange), all of which have half a unit of spin. According to this theory, the Dirac particles cannot interact directly with one another but only by exchanging energy quanta of the Yang–Mills field. The interactions between Yang–Mills fields and matter fields are identical for all particle types; only the Higgs field couples differently to the different matter fields. And only in this way is differentiation brought about between the various kinds of particle according to this new insight. By breaking the symmetry of the vacuum, the Higgs field could also give masses to the Yang–Mills quanta. But even the Higgs field is allowed to have only a limited number of interaction coefficients, so this model had only a small number of adjustable parameters: the masses of the quarks and leptons and a handful of 'mixing parameters'. The gravitational force, being excessively weak when acting between individual particles, could be included only to the extent that it acts classically.

The early versions of this model had other deficiencies. One of these was the remarkable absence of interactions due to the exchange of the neutral component, the Z boson, of the weak Yang–Mills quanta (the charged components being the W^+ and W^- bosons). These 'neutral current interactions' were detected for electrons and neutrinos in pivotal experiments at CERN in 1973 (Fig. 4). But they should also have caused strangeness-changing interactions among hadrons, and the existence of these was excluded by experimental observations. A possible remedy to this problem had already been proposed by Glashow, John Iliopoulos and Luciano Maiani in 1969, but this required a drastic revision of the model: the addition of a fourth quark, which was named charm.

The discovery of a series of new particles in 1974, beginning with the J/ψ particle at SLAC and at Brookhaven National Laboratory (the Alternating Gradient Synchrotron, in Upton, New York), marked a revolution of sorts. These new particles contained the elusive charm quark. Furthermore, their properties dramatically confirmed quantum chromodynamics and asymptotic freedom.

More details were then added. A rare type of transition observed in a special type of K meson called a K_L meson seemed to imply



PHYSICAL REVIEW VOLUME 96, NUMBER 1 OCTOBER 1, 1954

Conservation of Isotopic Spin and Isotopic Gauge Invariance*

C. N. YANG† AND R. L. MILLS
Brookhaven National Laboratory, Upton, New York
(Received June 28, 1954)

It is pointed out that the usual principle of invariance under isotopic spin rotation is not consistant with the concept of localized fields. The possibility is explored of having invariance under local isotopic spin rotations. This leads to formulating a principle of isotopic gauge invariance and the existence of a b field which has the same relation to the isotopic spin that the electromagnetic field has to the electric charge. The b field satisfies nonlinear differential equations. The quanta of the b field are particles with spin unity, isotopic spin unity, and electric charge ±e or zero.

INTRODUCTION

THE conservation of isotopic spin is ...cussed concept in recent years. ... isotopic spin parameter was first int... berg[1] in 1932 to describe the ... neutron and proton) ... stable even nuclei contain e... in 1937 Breit ...

Figure 3 | Yang–Mills gauge theory. The field equations introduced by Chen Ning Yang and Robert Mills in 1954 became the basis for the three forces of the standard model — electromagnetic, weak and strong. Image reproduced, with permission, from ref. 7.

breakdown of a symmetry called CP invariance. The most natural explanation of this was the existence of another pair of quarks, which were named top and bottom, because only the delicate interplay of at least six quarks with the Higgs field could give rise to the observed CP breakdown. Mathematical consistency of the scheme also required the existence of more leptons. The tau lepton, and its associated neutrino, were discovered and confirmed around 1978. The bottom quark (in 1977) and, finally, the top quark (in 1995) were also proved to exist.

Thus, a picture emerged of three generations, each containing two species of lepton and two species of quark. All of these particles interact with three types of Yang–Mills field and one Higgs field and, of course, with gravity. This picture was subsequently referred to as the standard model. In the 1970s, it was generally thought that this standard model would merely be a stepping stone. Amazingly, however, no improvements seemed necessary to explain the subsequent series of experimental observations. The standard model became a 'standard theory' — an accurate and realistic description of all of the particles and forces that could be detected.

One further detail did need to be added. The standard model was originally designed to accommodate only strictly massless neutrinos, but there was one anomaly — the neutrino flux from the Sun[5]. Pivotal observations announced in 1998, made using the Kamiokande detector, in Japan, showed that neutrinos can mix and therefore must have mass. Adding neutrino mass terms to the standard model was, however, only a minor repair and not totally unexpected, although it did add more parameters to the model. The earlier version had 20 fundamentally freely adjustable constants (parameters) in it; now, this number would need to be increased to at least 26.

Super theories

By the 1980s, it was understood that quantum field theories are perfect frameworks for the detailed modelling of all known particles. Indeed, if we require theories with only a limited number of elementary degrees of freedom, and thus a finite number of freely adjustable parameters, then it must be assumed that all forces are renormalizable. But, for all strong forces, the more stringent condition of asymptotic freedom is required. The only theories with these desired properties are theories in which Dirac particles interact exclusively with Yang–Mills fields and (where needed) with Higgs fields. This is now regarded as the answer to that problem of more than half a century ago — how to reconcile quantum mechanics with special relativity.

The mere fact, however, that these three Yang–Mills field systems are based on exactly the same general gauge principle, acting on the same sets of Dirac particles, has inspired many researchers not to stop here but to search for more common denominators. Can we find a completely unified field theory? Such theories have been sought before, notably by Einstein and by Werner Heisenberg in their later years, but their efforts were bound to fail because the Yang–Mills theories were then unknown. Now, it seems that we have the key to doing a much better job.

Indeed, we do have clues towards constructing a unified field theory. Despite its stunning successes, there are weaknesses in the standard model. Mathematically, the model is nearly, but not quite, perfect. Also, from a physics point of view, there are problematic features. One is the occurrence of gigantic differences in scale: some particles are extremely heavy, whereas others are extremely light. At distance scales that are short compared with the Compton wavelength of the heaviest particles — the cut-off scale below which field theories become important for these particles — there seems to be a crucial 'fine-tuning' among the effective couplings. And, most importantly, the quantum effects of the gravitational force are not included. These issues are the focus of new generations of theoretical proposals[6]. Might there be a new symmetry — a 'supersymmetry' — between Dirac particles and the force-carrying particles? Might particles turn out to be string-like rather than point-like? Or will a new generation of particle accelerators reveal that quarks and leptons are composites?

In the strongest possible terms, as theorists, we now urge our friends in experimental science to do whatever they can to obtain further information on the properties of nature's building blocks at the tiniest possible scales. In our business, this means reaching for the highest attainable energies: the Large Hadron Collider will make such a step. We can hardly wait. ■

Gerard 't Hooft is at the Institute for Theoretical Physics, Utrecht University and the Spinoza Institute, Post Office Box 80.195, 3508 TD Utrecht, The Netherlands.

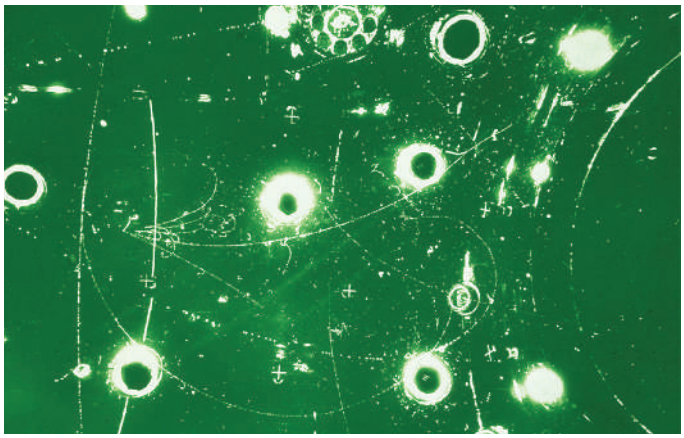


Figure 4 | The neutral current. An image from the heavy-liquid bubble chamber Gargamelle, at CERN, in 1973. The curling tracks reveal the interaction of a neutrino with a nucleon through the neutral current of *Z* exchange. Image reproduced with permission from CERN.

1. Pais, A. *Niels Bohr's Times, in Physics, Philosophy, and Polity* (Clarendon, Oxford, 1991).
2. Crease, R. P. & Mann, C. C. *The Second Creation: Makers of the Revolution in Twentieth-Century Physics* (Macmillan, New York, 1986).
3. Källén, G. *Elementary Particle Physics* (Addison-Wesley, Reading, Massachusetts, 1964).
4. Hoddeson, L., Brown, L. M., Riordan, M. & Dresden, M. (eds) *The Rise of the Standard Model: a History of Particle Physics from 1964 to 1979* (Cambridge Univ. Press, Cambridge, 1997).
5. Bahcall, J. N. *Neutrino Astrophysics* (Cambridge Univ. Press, Cambridge, 1989).
6. Ross, G. G. *Grand Unified Theories* (Perseus, Reading, Massachusetts, 2003).
7. Yang, C. N. & Mills, R. L. Conservation of isotopic spin and isotopic gauge variance. *Phys. Rev.* **96,** 191–195 (1954).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (g.thooft@phys.uu.nl).

NATURE|Vol 448|19 July 2007|doi:10.1038/nature06075

High-energy colliders and the rise of the standard model

Terry Wyatt[1]

Over the past quarter of a century, experiments at high-energy particle colliders have established the standard model as the precise theory of particle interactions up to the 100 GeV scale. A series of important experimental discoveries and measurements have filled in most of the missing pieces and tested the predictions of the standard model with great precision.

The standard model of particle physics describes the Universe as being composed of a rather small number of different types of elementary particle (see page 270) that interact in a small number of well-defined different ways.

Interactions among the elementary particles are represented by Feynman diagrams such as those in Fig. 1a. These show the annihilation of an electron–positron (e^+e^-) pair to produce a fermion–antifermion pair (such as a quark–antiquark or lepton–antilepton pair), and such interactions are examples of the 'electroweak' interaction, which is propagated by the photon, $W^{\pm}$ and Z bosons. All of the fermions participate in the electroweak interaction; certain 'self-interactions' among the photon, W and Z bosons may also take place.

Quarks, but not leptons, also participate in the strong interaction, which is propagated by gluons and described by the theory of quantum chromodynamics (QCD). Collectively, quarks and gluons are referred to as 'partons'. Quarks may carry one of three 'colours', which in the strong interaction are the analogue of charge; antiquarks carry the equivalent anticolour. A particular feature of the strong interaction is that coloured quarks cannot exist as free particles for more than about 10^{-24} s. The particles we observe in our detectors are hadrons — collections of quarks and/or antiquarks that have no net colour. There are two basic types of hadron: mesons contain a quark and an antiquark (of opposite colour); baryons contain three quarks (one of each colour). When a high-energy quark or gluon is produced, it is observed as a collimated 'jet' of hadrons.

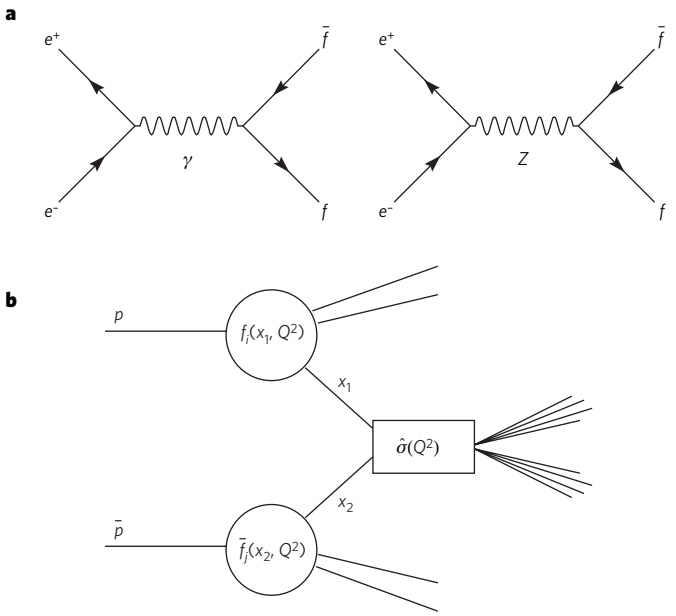


Figure 1 | Particle interactions. a, The lowest-order Feynman diagrams for the process $e^+e^- \rightarrow f\bar{f}$, where f is any elementary fermion (quark or lepton). **b**, Schematic view of a high-energy $p\bar{p}$ collision. Part b reproduced, with permission, from ref. 16.

The Higgs mechanism is introduced into the standard model to allow elementary particles to have non-zero masses, through their interaction with the Higgs field, while maintaining the gauge invariance of the model. A consequence of including the Higgs mechanism is that a massive, spin-zero Higgs boson is also predicted to exist.

If the mathematical structure of the standard model is taken as a given (although, of course, it represents a considerable amount of empirical input!), then all particle couplings are predicted in terms of a relatively small number of 'free' parameters that must be determined by experiments. For example, the strong interaction is determined by the value of a single coupling constant, denoted α_s. In the electroweak sector, the physically observed photon and Z boson arise from a linear superposition of two hypothetical particles: the W^0, the electrically neutral partner of the $W^{\pm}$, and another neutral boson, B^0 (of the so-called 'hypercharge' interaction). A rotation angle is defined between the W^0/B^0 and Z/photon, known as the electroweak mixing angle, θ_W, which describes the relative strengths of the electromagnetic and weak interaction. The interactions of the photon, Z and W are then determined by three free parameters. Logically, these can be thought of as the coupling constants of the weak and hypercharge interactions and the electroweak mixing angle. The masses of the W and Z bosons can also be predicted in terms of these parameters (with the photon and gluon required to be massless by gauge invariance). The masses of the 12 fermions and the Higgs boson are not predicted and thus represent additional free parameters that must be determined by experiment.

A particular feature of the electroweak interactions is that the couplings of the fermions to the W and Z depend on their handedness or helicity. The $W^{\pm}$ couples only to left-handed (negative-helicity) fermions and right-handed (positive-helicity) antifermions. The Z couples to both left- and right-handed fermions, but with a different coupling constant in each case.

In simple terms, the basic aims of particle physics are to find direct experimental evidence for each of the elementary particles and to make as precise as possible measurements of their various properties

[1]Particle Physics Group, School of Physics and Astronomy, University of Manchester, Manchester, UK.

(masses, coupling strengths and so on). Because the number of different experimental measurements that can be made is much larger than the number of free parameters in the standard model, we are dealing with an 'over-constrained' system. That is, our experimental measurements not only determine the values of the free parameters of the standard model, they also provide stringent tests of the consistency of the model's predictions.

Electrons versus protons

For more than a quarter of a century, the high-energy frontier of particle physics has been dominated by experiments performed at particle–antiparticle colliders. In these accelerators, beams of electrons and positrons, or protons (p) and antiprotons ($\bar{p}$), travel with equal and opposite momenta and collide head-on in the centre of the particle detectors.

Experiments at electron colliders have several advantages over those at proton colliders, which stem from the fact that electrons are elementary particles. When an e^+e^- pair annihilates, the initial state is well defined and, if the pair collide at equal and opposite momentum, the centre-of-mass energy of the system (E_{cm}) is equal to the sum of the beam energies. E_{cm} is the energy available to produce the final-state particles.

Electrons participate only in the electroweak interaction. This means that the total e^+e^- annihilation cross-section is small, so event rates in experiments are low, but essentially every annihilation event is 'interesting', and the observed events are relatively simple to analyse. Initial-state *bremsstrahlung* (radiation from the beam particles) can reduce the available centre-of-mass energy, but because this is a purely electromagnetic process it can be calculated with great precision, and it introduces no significant systematic uncertainties into the analysis of annihilation events.

The disadvantage of using electrons as beam particles is their small rest mass. When high-energy electrons are accelerated, they lose energy (producing synchrotron radiation), and that energy loss must be compensated by the machine's accelerating cavities. The energy radiated by a charged particle in performing a circular orbit of radius, R, is proportional to γ^4/R, where γ is the ratio of the particle's total energy to its rest mass, m_0c^2. Even though the world's largest particle accelerator, the Large Electron–Positron Collider (LEP), at CERN, had a circumference of 27 km, its maximum beam energy of around 104 GeV was limited by the fact that each particle radiated about 2 GeV per turn. By contrast, the large rest mass of the proton means that synchrotron energy loss is not a significant limiting factor for proton–antiproton colliders. For example, the world's highest energy collider at present is the Tevatron proton–antiproton collider, at Fermilab (Batavia, Illinois), which, with a circumference of only 6 km, achieves a beam energy of 1,000 GeV (or 1 TeV); the Large Hadron Collider (LHC), using two proton beams in the 27-km LEP tunnel, will achieve beam energies of 7 TeV.

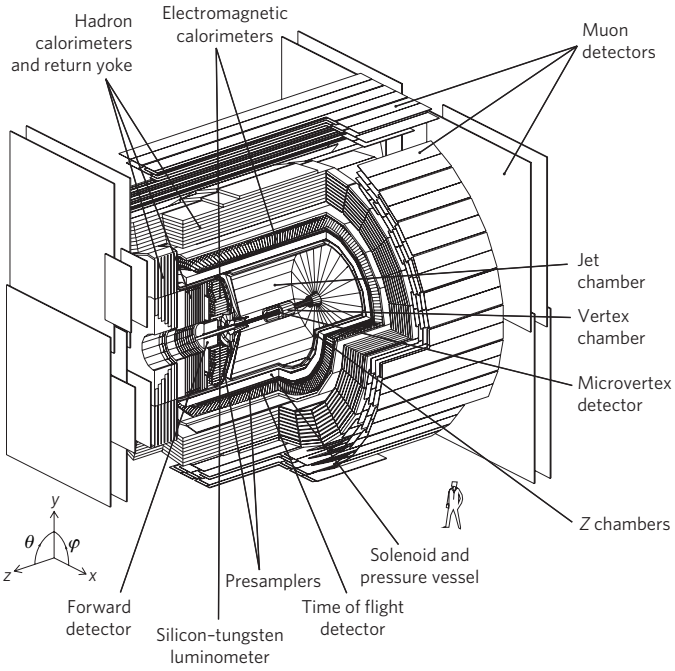


Figure 2 | The OPAL experiment at LEP. The typical, hermetic design of this detector comprises central track detectors inside a solenoid, calorimeters and — the outermost layers — muon detectors.

Although the beam energies of proton colliders may be much higher, for experiments at these colliders there are a number of challenges that stem from the fact that protons and antiprotons are strongly interacting, composite particles. A high-energy proton–antiproton collision is shown schematically in Fig. 1b. The highest energy collisions take place between a valence quark from the proton and an antiquark from the antiproton. These colliding partons carry fractions x_1 and x_2 of the momentum of the incoming proton and antiproton, respectively. The energy, Q, in the parton–parton centre-of-mass frame is given by $Q^2 = x_1x_2E_{cm}^2$. The probability of a proton containing a parton of type i at the appropriate values of x_1 and Q^2 is given by a 'parton distribution function' (PDF), $f_i(x_1, Q^2)$. The cross-section for the parton–parton collision to produce a given

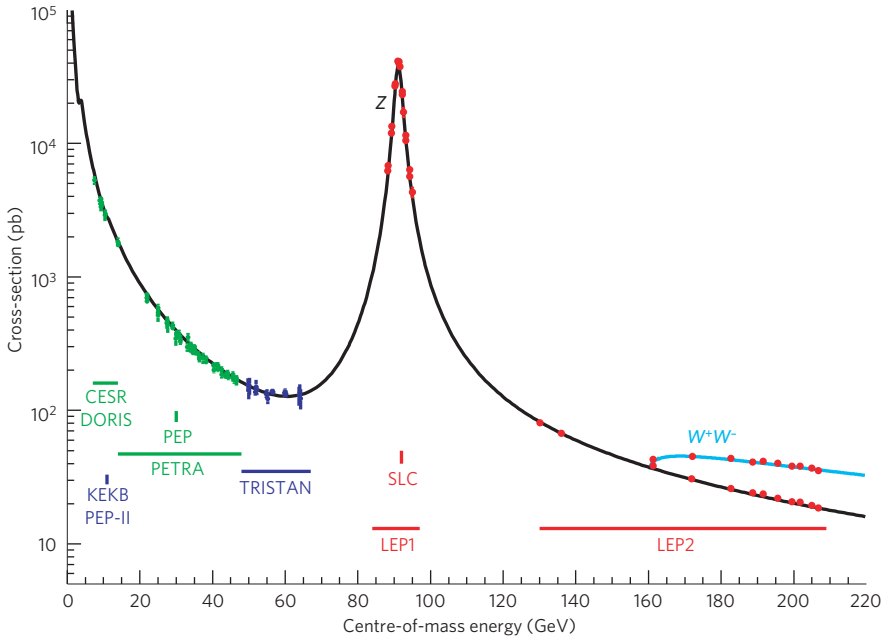


Figure 3 | The cross-section for e^+e^- annihilation to hadrons as a function of E_{cm}. The solid line is the prediction of the standard model, and the points are the experimental measurements. Also indicated are the energy ranges of various e^+e^- accelerators. (A cross-section of 1 pb = 10^{-40} m^2.) Figure reproduced, with permission, from ref. 12.

final state is denoted by $\hat{\sigma}(Q^2)$. To determine the cross-section, σ, for the proton–antiproton collision to produce this final state, we have to sum over all possible combinations of incoming partons and integrate over the momentum fractions x_1 and x_2:

$$\sigma = \sum_{i,j=q,\bar{q},g} \int dx_1 dx_2 f_i(x_1, Q^2) \cdot \bar{f}_j(x_2, Q^2) \cdot \hat{\sigma}(Q^2)$$

Therefore, the proton and antiproton beams, at a fixed beam energy, can be thought of as broadband beams of partons.

The total cross-section for proton–antiproton collisions at high energy is huge, and the event rate is consequently large — at the Tevatron, for example, about 10 collisions take place each time the bunches of protons and antiprotons meet and cross each other in the circular machine. Such bunch crossings take place 1.7 million times each second. But most of these collisions are rather uninteresting, because they result from a low momentum transfer between the proton and antiproton. Interesting processes, such as those containing W or Z bosons, are produced at a much lower rate and can be difficult to observe above the huge background.

Furthermore, the PDFs cannot be calculated from first principles in QCD. They can, however, be fixed by experimental measurements. A great deal of information on PDFs has come from the H1 and ZEUS experiments at the HERA collider, at DESY (Hamburg). At HERA, 27.5-GeV beams of electrons or positrons collide with a 920-GeV beam of protons, to produce 320 GeV in the centre-of-mass frame. The electrons and positrons provide a clean (electroweak-interaction) probe of the proton structure, and hence the PDFs, at these energies; the measured PDFs can then be extrapolated, using, for example, the so-called DGLAP evolution equations of QCD, to the much higher energies that are relevant at the Tevatron and the LHC.

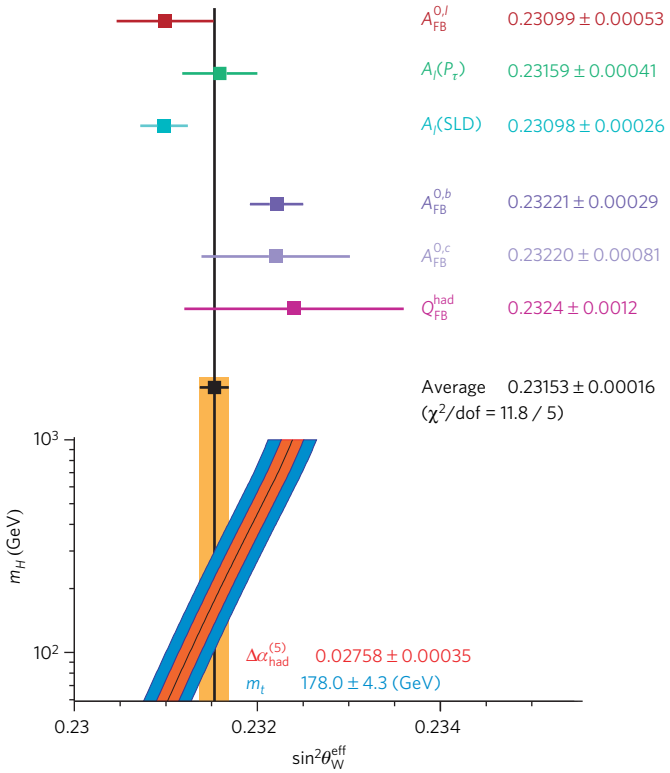


Figure 4 | Comparison of the effective electroweak mixing angle, $\sin^2\theta_W^{eff}$, derived from six classes of asymmetry measurements. These six classes are: A_{FB}, from leptonic final states at LEP; A_l (P_τ), derived from τ polarization measurements at LEP; A_l (SLD), derived from A_{LR} and from leptonic-final-state A_{LRFB} measurements at the SLC; and A_{FB}, from $b\bar{b}$ and $c\bar{c}$ final states at LEP. Q_{FB}^{had} is an average forward–backward charge asymmetry in hadronic events at LEP, without any attempt to distinguish individual quark flavours. Also shown is the standard-model prediction for $\sin^2\theta_W^{eff}$ as a function of m_H. The additional uncertainty of the standard-model prediction is parametric and dominated by the uncertainties in $\Delta\alpha_{had}^{(5)}(m_Z^2)$ (the correction for the effects of light-quark loops) and m_t, shown as bands. The total width of the band is the linear sum of these effects. Figure reproduced, with permission, from ref. 12. dof, degrees of freedom.

A further complication is that the initial-state partons have a high probability of radiating gluons before they collide. To some extent, this can be compensated by tuning Monte Carlo simulations of the collisions to those events that include leptonically decaying W and Z bosons (in which there is no complication from the possibility of final-state gluon *bremsstrahlung*). Nevertheless, the uncertainties associated with the lack of precise predictions for initial-state gluon *bremsstrahlung* represent a significant source of sytematic uncertainty in many analyses.

Proton and electron colliders are thus complementary: proton colliders offer the energy reach to make discoveries; electron colliders provide a cleaner experimental environment in which it is easier to make precise measurements.

Experiments at high-energy particle colliders typically share many common features, which are motivated by the requirements of the various measurements to be made. The basic aims are to detect with high efficiency each particle produced in the high-energy collision, to measure as accurately as possible its energy and momentum and to determine its particle type. No single detector type can achieve all of the above for all types of particle. Therefore, an experiment comprises a number of different detector systems, each of which has a specialized function. For example, at the centre of most experiments are detectors that measure the tracks produced by charged particles. Calorimeters are used for energy measurement, and muon systems are used for specific identification of those particles. An important feature of such detectors is their hermetic nature, which allows any apparent imbalance in the net transverse momentum of the visible particles to be ascribed to the production of weakly interacting particles, such as neutrinos. Fig. 2 shows, as an example, a cut-away view of the OPAL experiment at LEP, which is typical of detector design.

Discoveries and mounting evidence

By the late 1970s, the majority of the elementary fermions had been discovered. In particular, the discovery in the mid-1970s of the bottom quark and the tau lepton firmly established the existence of a third generation of fermions. However, there was only indirect evidence for the existence of two members of that generation: the top quark and the tau neutrino.

By contrast, among the elementary bosons only the photon had been observed directly as a physical particle. Although there was strong indirect evidence for the existence of the gluon, the first direct evidence came from the observations in 1979 by the JADE, Mark-J, TASSO and PLUTO experiments at the 30–35-GeV e^+e^- collider PETRA, at DESY. These experiments found events containing three hadronic jets, which correspond to the quark and antiquark produced in the e^+e^- collision, plus a gluon radiated from one of the quarks. The W and Z bosons were observed directly for the first time in 1983, by the UA1 and UA2 experiments[1–4] at the 560–640-GeV Super Proton Synchroton (SPS) proton–antiproton collider at CERN — a collider project that was conceived for the specific purpose of finding these particles and was rewarded with the 1984 Nobel Prize in Physics. The masses of the W and Z measured by the UA1 and UA2 experiments were found to be consistent with expectations, which was a beautiful confirmation of the standard model in electroweak interactions.

The scene was then set in 1989 for the 90-GeV e^+e^- colliders, LEP1 at CERN and SLC at the Stanford Linear Accelerator Center (SLAC; California). The ALEPH, DELPHI, L3 and OPAL experiments at LEP1 and the SLD experiment at the SLC performed measurements of Z production and decay that still today form the cornerstone of the precise tests of the electroweak standard model. Measurements of the Z mass elevated it to one of the most precisely known quantities within the standard model. Measurements of the total decay width of the Z (to all possible particle types) and the partial decay widths into each visible final state (that is, all final states except for $Z \rightarrow \nu\bar{\nu}$) allowed the number

of light neutrino species to be fixed at three. This observation effectively confirmed the existence of the tau neutrino as a distinct physical particle within the three-generation standard model and ruled out the existence of a fourth generation of fermions, unless the neutrino from that generation has a mass greater than half that of the Z. (Direct observation of the tau neutrino was finally reported[5] in 2001, in a different style of experiment using a proton beam directed at a fixed, tungsten target to produce neutrinos.)

Experimenters at all of the high-energy colliders since the days of PETRA had searched unsuccessfully for direct evidence for the existence of the top quark. These searches continued at 'Run I' of the 1.8 TeV Tevatron, which began in 1988. By 1994, the lower limit on the top quark mass from direct searches had reached[6,7] about 130 GeV. By that time, there was also considerable indirect evidence for the existence of the top quark. For example, measurements of the electroweak couplings of the bottom quark were consistent with the hypothesis that it formed one half of a pair of third-generation quarks within the standard model. Furthermore, fits to the precise electroweak data from LEP1 and SLC gave self-consistent results within the standard model only if the effects of a top quark with a mass of between about 155 and 195 GeV were included.

The top quark was directly observed for the first time in 1995, by the CDF and DØ collaborations[8,9] at the Tevatron. The first measurements gave its mass as 180 ± 15 GeV, consistent with the indirect determinations described above. This consistency represented a powerful confirmation of the electroweak standard model as an accurate picture of elementary particle physics.

In the second phase of the LEP programme, running between 1996 and 2000 with a vastly upgraded system of radio-frequency (RF) accelerating cavities, a maximum E_{cm} of nearly 209 GeV was reached. This allowed the production of W^+W^- pairs, enabling the ALEPH, DELPHI, L3 and OPAL experiments at LEP2 to measure the mass, m_W, and many other properties of the W boson.

In 2002, the second phase of running, Run II, began at the Tevatron. The accelerator complex was upgraded to deliver a slightly higher E_{cm} of 1.96 TeV and, more importantly, a greatly increased luminosity; the CDF and DØ detectors were also upgraded.

Now the most urgent question in particle physics (maybe in physics as a whole) is: where is the Higgs? Just as with the top quark, this question is being attacked on two fronts. Adding information from the direct measurements of the mass of the W boson and the top quark (from LEP2 and the Tevatron) to the precise electroweak measurements (from LEP1 and SLC) improves the precision with which the standard model can be tested. The overall fit gives self-consistent results only if the effects of a moderately light Higgs boson are included. Currently, a value for the Higgs mass of about 80 GeV is preferred, with an upper limit[10], at a 95% confidence level, of 144 GeV; further improvements to the mass measurements from the Tevatron may narrow the confidence interval. Direct searches for the Higgs boson were performed at LEP. The best available direct lower limit[11] on the Higgs mass is currently 114 GeV (95% confidence level) from the search for $e^+e^- \rightarrow ZH$ at LEP2. This already excludes a large part of the confidence interval allowed by the standard-model fit. Direct searches for Higgs production are currently the subject of intense effort at the Tevatron, and sensitivity to masses beyond the LEP2 limit is expected in the near future.

Precise tests of the standard model

A central part of the particle physics programme over the past quarter of a century has been to test the consistency of the standard model through precise measurement of many of its parameters. Precise theoretical calculations, implemented through computer codes of high technical precision, and a careful assessment of residual theoretical uncertainties are also essential elements in efforts to confront the standard model using precise data.

Let us return to the simple process shown in Fig. 1a, the annihilation of an e^+e^- pair to produce a fermion–antifermion pair through an electroweak interaction mediated by the photon or Z boson. Particles that

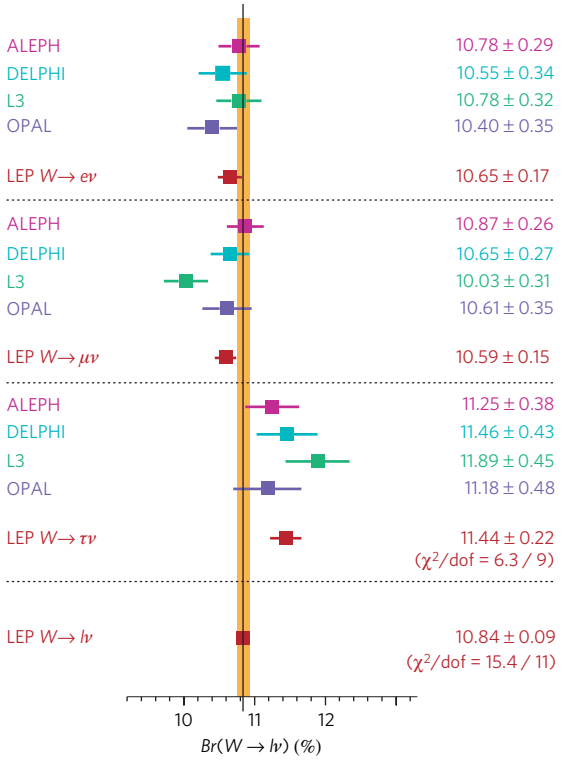


Figure 5 | Leptonic branching ratios. The measurements of the four LEP experiments of branching ratios for W decays to $e\nu$, $\mu\nu$ and $\tau\nu$ final states separately and for all lepton types combined. Figure reproduced, with permission, from ref. 14.

appear as internal lines in a Feynman diagram, such as the photon or Z in Fig. 1a, are 'virtual' particles — that is, they are not constrained to their 'physical' mass. However, the more virtual the particle becomes — the further away it is from its physical mass — the smaller the resultant amplitude for the process. Fig. 3 shows the cross-section for e^+e^- annihilation as a function of centre-of-mass energy, E_{cm}, based on data from several colliders including LEP and SLC. At low values of E_{cm}, the cross-section is dominated by the photon-exchange diagram (an exchanged Z would be highly virtual and the corresponding amplitude highly suppressed). With increasing E_{cm}, the cross-section falls as the exchanged photon becomes more and more virtual. At around 60 GeV, the amplitudes for photon and Z exchange are of comparable magnitude. As E_{cm} approaches the mass of the Z (91 GeV), the cross-section is dominated by the Z exchange diagram and reaches a peak, called the 'Z pole'.

The very large number of Z decays (around 20 million) collected by the experiments at LEP1 has allowed precise measurements of the couplings of the fermions to be made. The SLC delivered a much smaller number of Z decays (around 600,000) to the SLD experiment. However, the SLC delivered a longitudinally polarized e^- beam, which collided with an unpolarized e^+ beam, whereas at LEP both beams were unpolarized. The dependence on handedness of the fermion couplings has enabled SLD to make measurements, using polarized beams, that were in some respects competitive with and complementary to the measurements made at LEP1. (The results quoted in this section are all taken from ref. 12 unless explicitly stated otherwise.)

A number of important electroweak quantities have been determined from measurements around the Z pole at LEP1. The mass of the Z, m_Z, is related to the position of the peak in the cross-section, and total decay width of the Z, Γ_Z, is related to the width of the peak. The accuracy with which $m_Z = (91.1875 \pm 0.0021)$ GeV has been measured is limited by the accuracy with which the mean energy of the colliding beams is known over the entire data-taking period. Achieving such precision was a considerable challenge and resulted from a successful collaboration between physicists from both the LEP experiments and the accelerator. The energy

of a single circulating beam was determined to a high accuracy during dedicated calibrations, using the technique of resonant depolarization. However, such calibrations could be performed only every few days and gave the beam energy only at that specific point in time. The challenge was to propagate this precise knowledge of the beam energy over several days of accelerator running.

The circumference of the beam orbit is fixed by the frequency with which the RF accelerating cavities are excited. This frequency is very stable. The energy of the beams is then determined by the integral around the accelerator ring of the vertical component of the magnetic field experienced by the beams. This vertical magnetic field is produced mainly by the main 'bending' dipole magnets, but there is also a contribution from the large number of quadrupole magnets in the machine if the beam is not perfectly centred as it passes through them. If the position of the beam with respect to the quadrupoles changes over a period of hours or days this can affect the beam energy by a significant amount. Lunar tides, high rainfall in the nearby Jura mountains and changes in the water level of Lake Geneva all caused sufficient physical distortion of the accelerator (changing its radius by a few parts in 10^{-9}) to produce a measureable effect on the beam energy.

Erratic electric currents flowing in the accelerator beam pipe also affected the dipole fields over periods of many hours during which beams were circulating in the accelerator. Measurements of the spatial distribution of these currents around the ring established that they were produced by leakage currents from trains running on the Geneva-to-Bellegarde line. Understanding these various effects meant that a model could be developed to predict the beam energy as a function of time during data collection. Ultimately, residual uncertainties in the beam-energy calibration introduced systematic uncertainties of 0.0017 GeV in m_Z and 0.0012 GeV in Γ_Z, correlated among the four experiments.

The total decay width, $\Gamma_Z = (2.4952 \pm 0.0023)$ GeV, is given by the sum of the partial decay widths for each possible type of final-state fermion–antifermion pair. By measuring Γ_Z and the partial decay widths for each visible final state (quarks and charged leptons), the partial decay width to invisible final states (which in the standard model are neutrino–antineutrino pairs) can be determined. This number may be interpreted as a measurement of the number of types of light neutrino produced in Z decay, $N_\nu = 2.9840 \pm 0.0082$. This result requires the measurement of absolute cross-sections. These require a precise determination of the 'luminosity' of the accelerator, which is achieved by measuring the rate of low-angle electron–positron scattering. That the necessary precision of order 10^{-4} was achieved in these measurements represents a great success for theorists and experimentalists engaged in this joint project.

The rate of Z decays to quark–antiquark final states is enhanced by a factor related to α_s, the strong coupling constant, $(1 + \alpha_s^2/\pi + \ldots)$. Thus, a precise measurement of α_s can be made: $\alpha_s = 0.118 \pm 0.03$. This is in agreement with other precise determinations[13], such as those from event shapes (which are sensitive to the amount of final-state gluon radiation), and represents an important consistency test of QCD.

Asymmetries

Another class of electroweak measurement made at LEP1 and the SLC is of various asymmetries that are sensitive to the difference between the left- and right-handed couplings. One of the most sensitive of these electroweak measurements, and also one of the easiest to understand, is the so-called left–right asymmetry, A_{LR}. This is measured with polarized e^- beams at the SLC and is defined as:

$$A_{LR} = \frac{\sigma_L - \sigma_R}{\sigma_L + \sigma_R}$$

where σ_L (σ_R) is the cross-section for any given final state with a 100% left-hand (right-hand) polarized incoming electron beam. In practice, 100% polarization is not achievable, but it can be easily shown that if the magnitude of the (luminosity-weighted) average e^- beam polarization is $\langle P_e \rangle$ then the measured asymmetry, A_{LR}^{meas}, is given by $A_{LR}^{meas} = \langle P_e \rangle A_{LR}$. At the SLC, $\langle P_e \rangle = 70$–80% was regularly achieved.

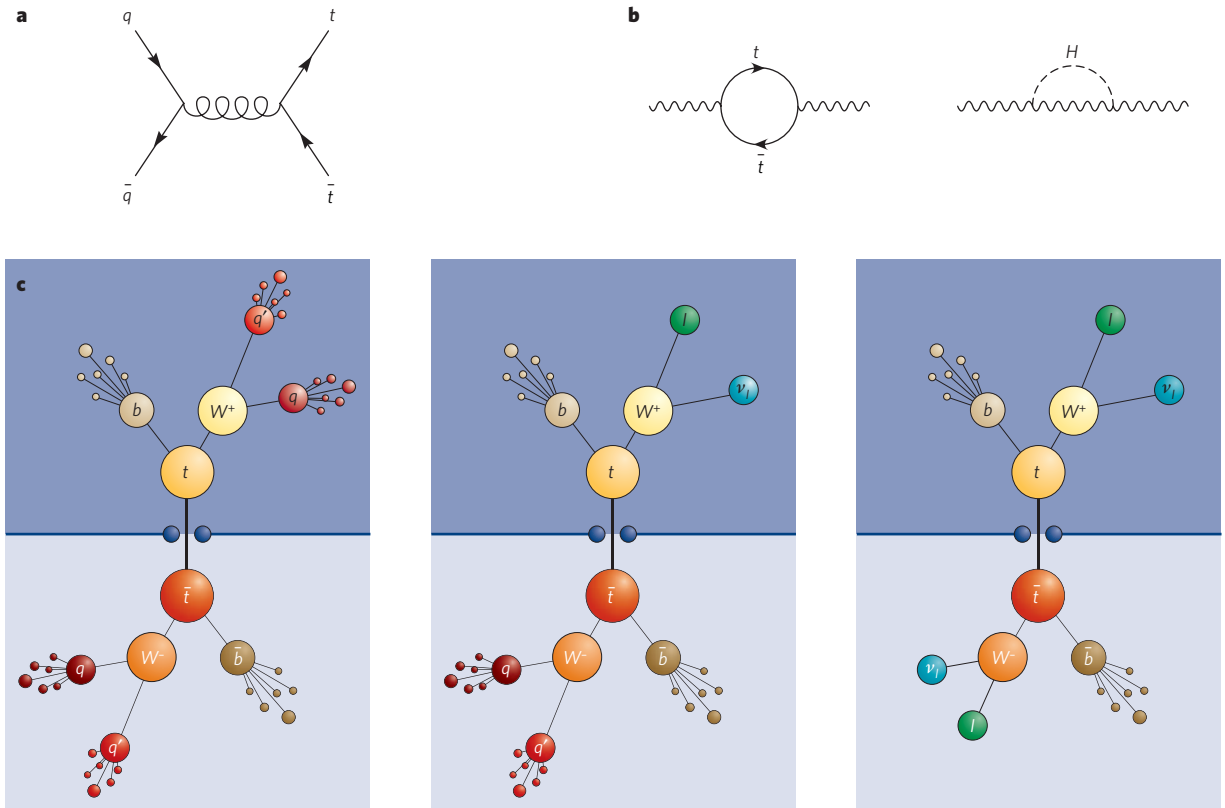


Figure 6 | Top-quark production, and virtual loops. a, The Feynman diagram for $q\bar{q}$ annihilation to produce a $t\bar{t}$ pair. **b**, Virtual loops involving t quarks and Higgs bosons. The left-hand diagram may modify a process involving the propagation of a photon or Z; the right-hand, the propagation of a W or Z. **c**, The possible event signatures for $t\bar{t}$ production. From left to right, panels show 'all-jets', 'lepton + jets' and 'di-lepton'.

The advantage of defining A_{LR} as above is that many factors — such as the dependence on the final-state couplings, acceptance of the detector, and so on — cancel in the ratio (as long as the experimental acceptance and $<P_e>$ are independent of the sign of the beam polarization). For such measurements at the Z pole, corrections (which are usually small) must be made to account for the photon-exchange diagram (Fig. 1a) and for the interference between the photon- and Z-exchange diagrams. In addition, a correction has to be applied for the fact that *bremsstrahlung* from the incoming e^+e^- results in an average annihilation centre-of-mass energy that is lower than the nominal E_{cm} of the colliding beams. Results are corrected to correspond to $E_{cm} = m_Z$, and these 'pole' cross-sections and asymmetries are therefore to be interpreted as corresponding to pure Z exchange at exactly $E_{cm} = m_Z$; they are sometimes denoted by adding the superscript '0' to the corresponding variable name, for example, A_{LR}^0.

As we have seen above, the fact that left- and right-handed e^- have different couplings from the Z produces an asymmetry between the annihilation cross-section for left- and right-hand-polarized incoming e^- beams. In addition, the difference between the left- and right-handed fermion couplings produces asymmetries in the angular distributions of the outgoing fermions. Consider an incoming e^- beam that is 100% left-hand polarized: angular-momentum conservation requires that this can annihilate only with the right-handed component of the incoming e^+ beam to produce Zs that are 100% polarized in the direction opposite to the incoming e^- beam. Angular-momentum conservation in the decay of the Z has the consequence that the preferred direction for the outgoing fermions to emerge is along the direction of the incoming e^- beam (the 'forward' direction) for left-handed fermions and in the opposite direction (the 'backward' direction) for right-handed fermions.

Using polarized electrons, as at the SLC, it is possible to define the 'left–right forward–backward' asymmetry,

$$A_{LRFB} \equiv \frac{(\sigma_F - \sigma_B)_L - (\sigma_F - \sigma_B)_R}{(\sigma_F + \sigma_B)_L + (\sigma_F + \sigma_B)_R}$$

As before, the measured asymmetry, A_{LRFB}^{meas}, is given by $A_{LRFB}^{meas} = <P_e> A_{LRFB}$.

At LEP, the e^- and e^+ beams were unpolarized. That is, there were equal numbers of left- and right-handed incoming beam particles. Nevertheless, the fact that left- and right-handed e^- have different couplings to the Z produces an asymmetry between the numbers of left- and right-hand incoming e^- that annihilate. Thus, the produced Zs are partially polarized along the direction of the incoming beams, and the difference between the left- and right-handed fermion couplings produces a forward–backward asymmetry, A_{FB}, in the angular distributions of the outgoing fermions, which is given by:

$$A_{FB} \equiv \frac{(\sigma_F - \sigma_B)}{(\sigma_F + \sigma_B)}$$

The forward–backward asymmetry with unpolarized beams, A_{FB}, mixes the couplings of the initial- and final-state particles. This makes A_{FB} intrinsically a less sensitive measure of the electroweak mixing angle, θ_W, (in the form $\sin^2\theta_W$) than measurements possible with polarized beams. However, the much larger samples of Zs available at the LEP experiments compensate for this lack of intrinsic sensitivity.

To measure A_{LRFB} and A_{FB}, it is necessary to isolate a sample of Z decays to a particular fermion type and to distinguish the fermion from the antifermion. In the case of Z decays to charged leptons this is fairly straightforward: events containing a high-momentum e^-e^+, $\mu^-\mu^+$ or $\tau^-\tau^+$ pair may be readily distinguished from one another and from other backgrounds; the electric charge distinguishes the lepton from the antilepton. In the case of Z decays to quarks, precise measurements of A_{FB} are only really possible in the $c\bar{c}$ and $b\bar{b}$ final states.

In most cases it is not possible to determine the handedness of the final-state particles (hence observables are usually summed over this quantity). The one exception is for final-state tau leptons, where the momenta of the observed tau decay products are correlated with the handedness of the produced tau.

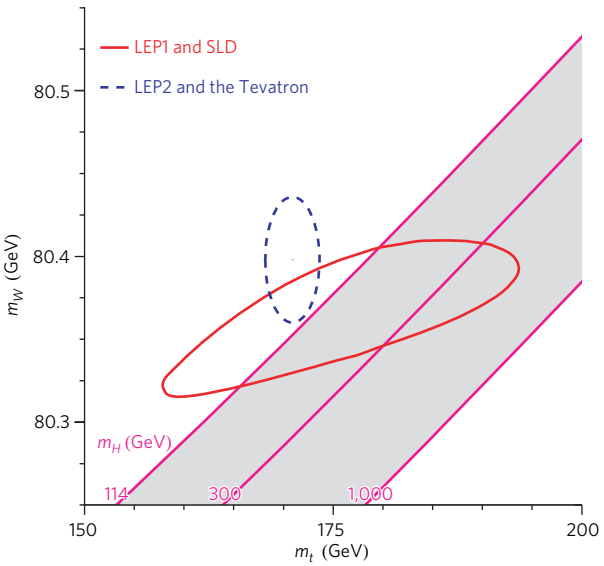


Figure 7 | Contours at 68% confidence level showing the direct (LEP2 and the Tevatron) and indirect (LEP1 and SLC) measurements of m_W and m_t. The shaded band shows the predictions of the standard model for various values of m_H. Figure reproduced, with permission, from ref. 10.

All of the asymmetry measurements discussed are sensitive to the difference between the left- and right-handed fermion couplings and thus to the value of $\sin^2\theta_W$. The degree to which the different classes of asymmetry measurements yield consistent values of $\sin^2\theta_W$ — as illustrated in Fig. 4 —represents an important consistency check of the standard model.

Consistency of the standard model

In W^+W^- events at LEP2, the value of m_W is obtained by directly reconstructing the invariant mass of the pair of particles produced in the W decay. In principle, the two final states with high branching ratios — $q\bar{q}l\bar{\nu}$ and $q\bar{q}q\bar{q}$ — give similar statistical sensitivity. However, in the $q\bar{q}q\bar{q}$ channel, uncertainties associated with strong interactions and Bose–Einstein correlations between the products of the two hadronically decaying Ws render the measurement of m_W in this channel less precise. The combination of results[14] from the four LEP experiments yields $m_W = (80.376 \pm 0.033)$ GeV. Other properties of the W (such as the branching ratios shown in Fig. 5) were measured[14] at LEP2.

At the Tevatron, only the leptonic decays $W \to e\nu$ and $W \to \mu\nu$ can be used to measure m_W. CDF has produced the first preliminary measurement of m_W using the Run II data accumulated so far, and it has an uncertainty to match that of a single LEP experiment. Including data from Run I, the Tevatron average[10] is $m_W = (80.429 \pm 0.039)$ GeV. Combining the LEP and Tevatron values gives the 'world average'[10] as $m_W = (80.398 \pm 0.025)$ GeV.

The most important process for producing top quarks in $p\bar{p}$ collisions is shown in Fig. 6a. The dominant decay of the top quark is $t \to Wb$ and possible signatures of $t\bar{t}$ production are shown schematically in Fig. 6c. If one W decays leptonically and one W decays hadronically, a final state is produced containing a high-transverse-momentum lepton, missing transverse momentum (due to the undetected neutrino) and four high-transverse-momentum jets. This occurs in about 46% of $t\bar{t}$ pairs produced, and this so-called 'lepton + jets' channel yields the most precise measurement of m_t. The combination[15] of CDF and DØ measurements gives $m_t = (170.9 \pm 1.8)$ GeV. This precision of around 1% makes m_t by far the most precisely known quark mass. The ultimate precision expected for the Tevatron measurements is around 20 MeV for m_W and around 1 GeV on m_t; to equal such precision at the LHC will take much time and concerted effort.

It is interesting to understand how experiments can produce evidence for the existence of a particle, and even constrain its mass and couplings,

even though they have insufficient energy to produce the particle directly. The indirect effects of the top quark and the Higgs boson may be observed at LEP/SLC because of the existence of processes such as those shown in Fig. 6b. The possibility of such 'radiative corrections' modifies the simple 'lowest-order' picture of e^+e^- annihilation in Fig. 1a, and experimentally observable effects become sensitive to the masses and couplings of virtual particles in such loops. For example, it is usual to consider $\sin^2\theta_W^{\text{eff}}$, an 'effective' parameter that absorbs the effect of the radiative corrections but allows the basic form of the coupling equations involving $\sin^2\theta_W$ to stay the same. The correction to $\sin^2\theta_W$ can be calculated in the standard model; it depends on the square of the top quark mass, m_t, but only logarithmically on the Higgs mass, m_H. An illustration of these effects is given in the lower half of Fig. 4, in which the experimentally measured value of $\sin^2\theta_W^{\text{eff}}$ is compared with the prediction of the standard model as a function of m_H.

The contours in Fig. 7 show the world-average direct measurements of m_W and m_t compared with the indirect values of those quantities extracted from the standard-model fit to the LEP and SLC data. The shaded band shows the predictions of the standard model for various values of m_H. The fact that the direct and indirect values of m_W and m_t agree is a triumph of the standard model.

An even more stringent test of the consistency of the standard-model fit to all available high-energy electroweak data is shown in Fig. 8: each measured quantity is compared with its value obtained from the fit. The largest single deviation is seen for $A_{FB}^{0,b}$ (the forward–backward asymmetry for Z decays to bottom quarks) measured at LEP1, but, particularly given the number of measurements considered, a discrepancy of 2.8 standard deviations in one of them does not meet the threshold required for claiming a significant departure from the standard model.

The increased samples of $t\bar{t}$ events available at Run II have allowed measurements of the cross-section for $t\bar{t}$ production and of t quark properties, such as spin, electric charge and decay branching ratios, that are

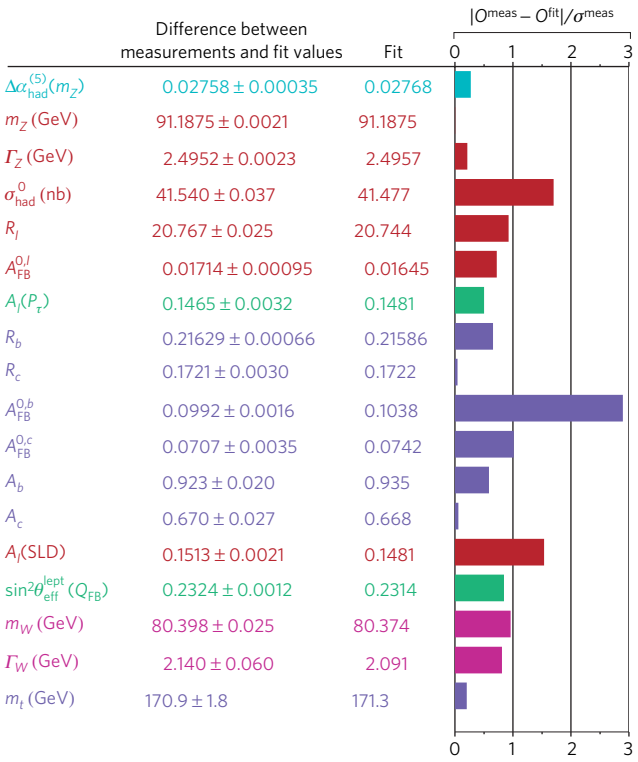


Figure 8 | A test of the consistency of the standard-model fit to all available high-energy electroweak precise data. Each measured observable (O^{meas}) quantity is compared with the value obtained from the fit (O^{fit}). Also shown graphically is the difference between measurement and fit values in number of standard deviations. Colours indicate groups of similar variables. Figure reproduced, with permission, from ref. 10. For full definitions of each quantity, see ref. 10.

consistent with those expected in the standard model[16]. The Tevatron experiments are also detecting processes with ever smaller cross-sections— which bodes well for developing the sensitivity of the searches for the Higgs boson at this collider. The CDF experiment has detected the associated production of WZ pairs[17] and has found the first evidence at a hadron collider for the production of ZZ pairs[18]; the DØ experiment has found the first evidence for electroweak production of single top quarks[19], enabling the first direct determination of the $t \to Wb$ coupling.

The years ahead

The next few years will be an exciting time in experimental particle physics, with first collisions at the 14 TeV proton–proton collider, the LHC, scheduled for 2008. Until then, as the world's current highest-energy collider, the Tevatron has a monopoly on direct searches for new physics at a high-mass scale and can perform the most stringent tests of the point-like nature of the fundamental particles.

The Tevatron will run at least until late 2009; its mantle will not pass to the LHC overnight. Except for a few special cases that could produce the most spectacular, unmistakable signatures, it will take time to understand and calibrate the LHC accelerator and detectors.

It is hard to imagine that new physics beyond the standard model will not be found at the LHC. What form that new physics will take is harder to imagine. We know from the past 30 years' work that all theories predicting any observable effects beyond the predictions of the standard model were quickly disposed of by experiment. This means that no matter what is to come, the standard model will remain at least an extremely accurate 'approximation' to the physics of elementary particles at scales up to a few hundred GeV. ■

1. Arnison, G. *et al.* (UA1 Collaboration). Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s} = 540$ GeV. *Phys. Lett. B* **122,** 103–116 (1983).
2. Banner, M. *et al.* (UA2 Collaboration). Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN $\bar{p}p$ collider. *Phys. Lett. B* **122,** 476–485 (1983).
3. Arnison, G. *et al.* (UA1 Collaboration). Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider. *Phys. Lett. B* **126,** 398–410 (1983).
4. Bagnaia, P. *et al.* (UA2 Collaboration). Evidence for $Z^0 \to e^+e^-$ at the CERN p collider. *Phys. Lett. B* **129,** 130–140 (1983).
5. Kodama, K. *et al.* (DONUT Collaboration). Observation of tau neutrino interactions. *Phys. Lett. B* **504,** 218–224 (2001).
6. Abachi, S. *et al.* (DØ Collaboration). Top quark search with the DØ 1992–1993 data sample. *Phys. Rev. D* **52,** 4877–4919 (1995).
7. Abachi, S. *et al.* (DØ Collaboration). Search for the top quark in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV *Phys. Rev. Lett.* **72,** 2138–2142 (1994).
8. Abe, F. *et al.* (CDF Collaboration). Observation of top quark production in $p\bar{p}$ collisions with the collider detector at Fermilab. *Phys. Rev. Lett.* **74,** 2626–2631 (1995).
9. Abachi, S. *et al.* (DØ Collaboration). Observation of the top quark. *Phys. Rev. Lett.* **74,** 2632–2637 (1995).
10. LEP Electroweak Working Group. *LEP Electroweak Working Group*. <http://lepewwg.web.cern.ch/LEPEWWG/> (2007).
11. ALEPH Collaboration, DELPHI Collaborations, L3 Collaboration, OPAL Collaboration and The LEP Working Group for Higgs Boson Searches. Search for the standard model Higgs boson at LEP. *Phys. Lett. B* **565,** 61–75 (2003).
12. ALEPH, DELPHI, L3, OPAL, SLD collaborations, LEP Electroweak Working Group, the SLD Electroweak and Heavy Flavour Groups. Precision electroweak measurements on the Z resonance. *Phys. Rep.* **427,** 257–454 (2006).
13. Jones, R. W. L. Final α_s combinations from the LEP QCD working group. *Nucl. Phys. B Proc.* (suppl.) **152,** 15–22 (2006).
14. The LEP Collaborations: ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, the LEP Electroweak Working Group. A combination of preliminary electroweak measurements and constraints on the standard model. Preprint at <http://arxiv.org/abs/hep-ex/0612034> (2006).
15. Tevatron Electroweak Working Group (for the CDF and DØ Collaborations). A combination of CDF and DØ results on the mass of the top quark. Preprint at <http://arxiv.org/abs/hep-ex/0703034v1> (2007).
16. Quadt, A. Top quark physics at hadron colliders. *Eur. Phys. J. C* **48,** 835–1000 (2006)
17. Abulencia, A. *et al.* (CDF Collaboration). Observation of WZ production. Preprint at <http://arxiv.org/abs/hep-ex/0702027>.
18. CDF Collaboration. Evidence for ZZ production in $p\bar{p}$ at $\sqrt{s} = 1.96$ TeV. CDF Note 8775 (preliminary). <http://fcdfwww.fnal.gov/physics/ewk/2007/ZZ/ZZ_comb_public_note.ps>.
19. Abazov, V. M. *et al.* (DØ Collaboration). Evidence for production of single top quarks and first direct measurement of |Vtb|. *Phys. Rev. Lett.* **98,** 181802 (2007).

Author information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to T.W. (twyatt@fnal.gov; Terry.Wyatt@manchester.ac.uk).

How the LHC came to be

Chris Llewellyn Smith

Approval of a project the size of the Large Hadron Collider is an exercise in politics and high finance.

The idea of following CERN's Large Electron–Positron Collider (LEP) with a Large Hadron Collider (LHC), housed in the same tunnel, dates back at least to 1977, only two years after LEP itself was conceived. The importance of not compromising the energy of an eventual LHC was one of the arguments for insisting on a relatively long tunnel in the discussions that led to the approval of LEP in 1981.

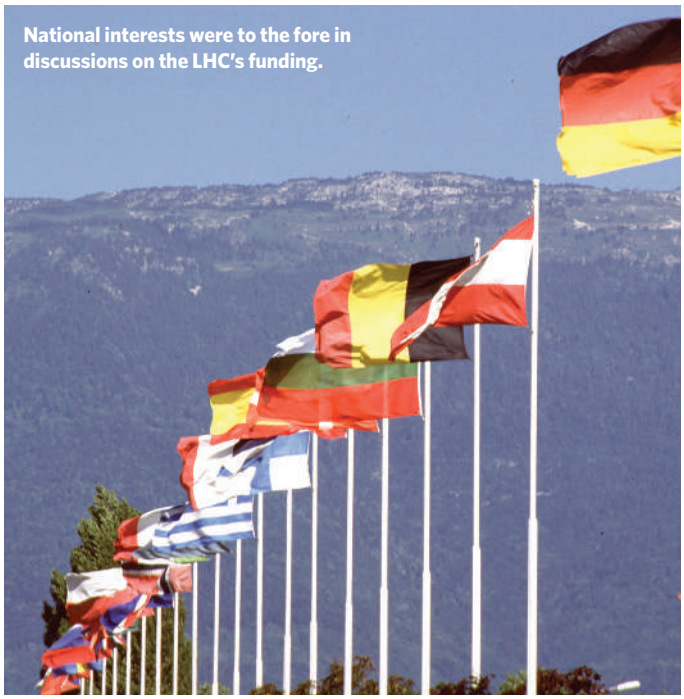
Early discussions of the LHC were dominated by sometimes acrimonious competition and comparisons with the proposed 40 teraelectronvolt (TeV) Superconducting Super Collider (SSC) in the United States. Serious work on the SSC was kick-started by American reactions to the discovery of the carriers of the weak force, the $W^{\pm}$ and Z bosons, at CERN in 1983. CERN's discovery was greeted by a *New York Times* editorial entitled "Europe 3, US Not Even Z-Zero", and a call from the President's science adviser for the United States to "regain leadership" in high-energy physics.

Viewed from Europe, this was provocative. 40 TeV is more than twice the energy that could possibly be reached by a hadron collider installed in the LEP tunnel, and many Europeans suspected that this was why 40 TeV was chosen. Furthermore, CERN's leadership favoured the next really large accelerator being an inter-regional facility. There was agreement on the vital importance of being able to explore new phenomena of up to 1 TeV — the energy below which the Higgs boson, or whatever else generates the mass of all particles, should be discovered. But how much better this could be done at the SSC than at the lower energy LHC was hotly debated. The energy domain that can be explored by a hadron collider is less than that of the accelerated particles, which is shared between their constituents. However, the much higher intensity (or luminosity, in the terminology of particle physics) planned for the LHC could in principle compensate for it having lower energy than the SSC — with denser bunches of accelerated particles there is a greater chance of collisions between constituents with large fractions of their parents' energy — although another decade of intensive research and development was needed to establish that experiments are possible at such high luminosity.

"Not everyone was convinced that the collider was justified, or would ever be funded"

The extreme European view was that the SSC was irresponsible as it would cost five times as much as the LHC without providing much more physics. A global plan, on the other hand, would provide complementary facilities — a large linear electron–positron collider in the United States and the LHC at CERN, which would use the LEP tunnel and other existing infrastructure — and would therefore be much cheaper than the SSC. A typical American response was to refute the claim that the LHC could do much the same physics for one-fifth of the cost, and to meet with scepticism any concern for American taxpayers. Meanwhile, senior Japanese physicists who argued that the SSC should be international were told that it was a national facility. They did not forget this when the United States later asked Japan to contribute US$2 billion.

I thought the extreme European position was unrealistic. The United States wanted a new project that could reach 1 TeV as soon as possible and a large hadron collider was the only realistic option at the time. The technology was not then available to build a 1 TeV linear collider: the 0.1 TeV linear collider at Stanford, the world's first, had not been built, and even the International Linear Collider now being proposed will initially reach only 0.5 TeV. On the other hand, my experience as adviser to the Kendrew Committee, which was then considering whether the United Kingdom should remain in CERN, had made me acutely conscious of growing pressure on funding for particle physics. I doubted that the SSC — which seemed profligate to me, despite its enormous potential — would ever be funded.



National interests were to the fore in discussions on the LHC's funding.

CERN

In fact, the SSC was endorsed by President Ronald Reagan in January 1987, with a price tag of $4.4 billion. In May 1990, when the cost had risen to $7.9 billion, the House of Representatives voted to limit the federal contribution to $5 billion, with the rest to be found from the state of Texas ($1 billion) — the proposed site of the SSC — and from overseas (where none was found). The SSC was defeated in the House in June 1992, but later revived by the Senate; this happened again in June 1993, by which time the General Office of Accounting estimated the cost as $11 billion. It was cancelled in October 1993.

The right machine for the future

Meanwhile at CERN, research and development started on the very demanding LHC magnets in 1988. This was recommended by a Long Range Planning Group chaired by Carlo Rubbia, who shared the 1984 Nobel Prize in physics for the discovery of the W and Z bosons, and who became Director-General of CERN in 1989. Rubbia argued that the LHC would provide healthy competition for the SSC at a relatively modest cost and that it would be more versatile and bring important

additional new physics. As well as accelerating protons, it would be able to accelerate heavy ions to world-beating energies at little extra cost, and LHC protons could be collided with electrons in LEP at much higher energy than in the Hadron Electron Ring Accelerator (HERA) then being built in Hamburg (this option was abandoned in 1995 when it was decided that, after it was eventually closed, LEP should be removed to make it easier to install the LHC).

Rubbia's powerful advocacy and a series of workshops built up pan-European enthusiasm for the LHC, although not everyone was convinced that the collider was justified, or would ever be funded, in parallel with the SSC. I swallowed any doubts and supported the LHC as a member (1986–92) and chairman (1990–92) of CERN's scientific policy committee. I thought the versatility argument was good, and that the LHC should be supported — at the very least as an insurance policy in case the SSC ran into trouble.

In December 1991, the CERN council adopted a resolution that recognized the LHC as "the right machine for the advance of the subject and the future of CERN" and asked Rubbia to come forward with a complete proposal before the end of 1993. I was due to succeed Rubbia as CERN's Director-General at the beginning of 1994, and in early May 1993 he handed me the responsibility of preparing and presenting the LHC proposal. The outlook was not encouraging: a new, detailed costing was significantly bigger than previous estimates; the personnel requested by CERN group leaders to build the LHC would have required a 20% increase in staff; attitudes to high-energy physics were hardening in several CERN member states; and the CERN council had just agreed to a temporary reduction in Germany's contribution on the grounds that reunification was proving very costly.

Bid for approval

Over the summer and autumn of 1993, Lyn Evans, by then nominated as LHC director, proposed several modifications to the design that reduced the cost, and with the help of many others I identified reductions in the rest of the CERN programme that would free up money and manpower. Costing was difficult as this was before most of the research and development for the LHC had been completed — for instance, the first full-length dipole magnet was not tested until December 1993. It was also before approval of the experimental programme, which became more ambitious after a large influx of American researchers joined proposals for LHC experiments after the SSC was cancelled. Our 1993 costing therefore underestimated the eventual specification and cost of the underground areas that were to house the experiments.

The plan I presented to the CERN council in December 1993 foresaw LHC construction, with commissioning in 2002, on the basis of a humped budget, with full compensation for inflation of materials costs. The hump was to come from a mixture, still to be defined, of a general budget increase, additional voluntary contributions from some member states and contributions from non-member states. The plan was generally well received, although it was clear that Germany and the United Kingdom were very unlikely to agree a budget increase, and we were asked to come back with proposals to reduce costs further and indications of how much non-member states might be willing to contribute.

We developed proposals to delay LHC commissioning until 2003 or 2004, stage the construction of the detectors and, while maintaining priority for LEP and the very ambitious LEP upgrade (the first phase of which was not complete), reduce other parts of the CERN programme to a bare minimum over the coming years, with complete closure for one year. In June 1994 we requested approval of the LHC from the CERN Council. The date of commissioning was to be decided later, depending on what voluntary and non-member state contributions were obtained. Encouragingly, a US panel had by then recommended that "the government should declare its intention to join other nations in constructing the LHC" (although the suggested contribution was disappointingly small), and positive signals had been received from Japan, Russia and India. Seventeen member states voted to approve the LHC. The vote was left open, however, because the other two member states — Germany and the United Kingdom — would not accept the proposed budgetary conditions, and demanded substantial additional voluntary contributions from the host states (Switzerland and France), who, they considered, gained disproportionate benefits from CERN.

The missing-magnet machine

Over the next six months, difficult discussions ensued between CERN, the host states, and Germany and Britain (at one point, the Director General of the UK Research Councils, John Cadogan, told me I would be "staring into the abyss" if we could not reduce the cost of the LHC). Some movement on the part of France and Switzerland was beginning to ease the position when Germany and Britain announced that they could only approve the LHC under a planning assumption of 2% inflation to be compensated by 1% indexation — in other words, a 1% annual budget reduction in real terms — and continuation of the German rebate for

Even before the 27-kilometre tunnel for the Large Electron–Positron collider was built, thoughts were turning to its successor, the Large Hadron Collider.

CERN

some years. These conditions seemingly made it impossible to launch the LHC while upgrading and exploiting LEP, and maintaining CERN's small programme of excellent fixed-target experiments.

Our response was to propose keeping down the annual cost by building a 'missing-magnet machine', in which a third of the dipole magnets would be omitted in a first phase, thereby saving some 300 million Swiss francs (US$240 million). The machine would have operated at two-thirds of full energy for some years, before the remaining magnets were installed. Although the final cost would have been more, and results from phase two would have eclipsed the physics from phase one, the two-stage LHC would nevertheless have been a world-beating facility — and it was the only option available.

I was asked whether the two-stage LHC would be viable under the proposed budget conditions. My first reply amounted to a "yes, but…". It was made clear that qualifications would prevent approval. So I took a deep breath and replied yes, adding that the conditions would be acceptable if accompanied by an assurance that any contributions from non-member states would be used to speed up and improve the project, not to allow reductions in the member states' contributions. The CERN council explicitly gave this assurance, and on 16 December 1994 approved the LHC, on the basis of a two-stage construction plan, to be reviewed in 1997 in light of what contributions had been obtained from non-member states, and of the budget conditions required by Germany and Britain, including the continuation of the German rebate. Generous French and Swiss offers to make in-kind contributions and to increase their contributions by 2% a year were crucial factors.

Reassured by a letter from the then British Minister of Science, David Hunt, which described the conditions as "realistic, fair and sustainable", it seemed that all that remained was to try to identify further internal cutbacks, so as to turn our reply to the question of viability into a genuinely unqualified yes, and to seek contributions from non-member states. This we did. Negotiations with Japan — which made the first-ever substantial contribution of a non-member state to a CERN accelerator in June 1995 — and with Russia, India, Canada and the United States went well. By the middle of 1996, we were becoming confident that single-stage construction would be possible, and raised with the council the possibility of bringing forward the 1997 review to December 1996.

Then, in July 1996, out of the blue, the German government announced that, to help ease the financial burden of reunification, it intended to reduce all international science subscriptions. A particularly large cut (8.5% for two years and 9.3% thereafter) was proposed for CERN, despite the fact that Germany was already enjoying a 'reunification rebate'. The possibility of limiting the reduction to just the German contribution was scuppered when the UK government announced that, the minister's letter notwithstanding, it had always seen the 1997 review as another opportunity to look for reductions in the CERN budget. It called for "the largest possible reduction" claiming that this could be achieved "without damaging CERN's scientific mission or endangering the LHC". The particle-physics community in the United Kingdom was reluctant to challenge this assertion, as they were told that reductions in the CERN budget would be their source of funding for participation in LHC experiments.

Forwards through deficit

CERN's public response to the proposed budget cut was muted by fear of shaking the confidence of the non-member states, who had been assured that their contributions would not be used to allow reductions in the member states' contributions. The United States, which had repeatedly asked for reassurances on the viability and sustainability of CERN's planning and funding, was a particular worry. A US contribution to the LHC, significantly larger than suggested by the 1994 panel, had just been negotiated, but a formal agreement had not yet been drafted, let alone signed.

The subsequent discussions were rough. At one stage, Germany threatened to leave CERN if its demands were not met exactly, and even prepared a letter of withdrawal that leaked to the press. There were suggestions that the United Kingdom might also use the threat of withdrawal as a negotiating tactic. I repeatedly told the CERN council that cuts of the magnitude proposed would destroy the LHC's viability, although — as inflation had been zero since 1994, thanks to the strength of the Swiss franc — we had built up a reserve of around 2% in the budget, which could be sacrificed without making matters worse than they had been at the end of 1994. Drawing attention to this reserve was perhaps a tactical error, as it seemed to make a 9% reduction somewhat less out of reach. Given the United Kingdom and Germany's determination, however, I don't think it made any difference to the final outcome.

The crunch came in October when with Horst Wenninger, who played a central role in putting together the LHC proposal, I met the German minister with responsibility for CERN, Jürgen Rüttgers. We explained that the LHC could not survive the proposed budget cut. Rüttgers was uncompromising, but finally asked whether there was any

conceivable way to avoid this conclusion. We replied that it could be avoided if CERN were allowed to take out loans, an idea that Germany had previously vetoed: deficit financing would be extremely risky, but would be acceptable if accompanied by clear acknowledgement of the risks by the CERN council and approval of single-stage construction, which was not only desirable scientifically, but necessary to ensure the contributions required from the non-member states.

A sting in the tail

During the next meeting of delegations to CERN, Germany declared that "a greater degree of risk would inevitably have to accompany the LHC". Others, while accepting deficit financing, acknowledged the risks in similar or stronger terms, and single-stage construction of the LHC, with completion foreseen in 2005, was approved on 20 December 1997. After intensive lobbying of other member states, the accompanying budget reduction was marginally smaller than requested by Germany. The CERN council also imposed a one-year 2% 'crisis levy' on the salaries of CERN staff, even though they had hardly risen since 1993, while further efficiency savings and economies were sought.

We duly got on with this job, but there was a major distraction and sting in the tail. The new chairman of the US House of Representative's Science Committee, James Sensenbrenner, who was suspicious of international projects as a result of his experience with the International Space Station, declared that the proposed agreement between the US Department of Energy and CERN was unsatisfactory. I believed that without the United States, the hard-won European agreement to build the LHC might unravel, whereas their involvement would make it secure. Some very anxious months ensued. In the end it turned out that Sensenbrenner was satisfied with modifications that strengthened the United States' protection against unforeseen events without changing the magnitude of its contribution, and a US–CERN agreement was finally signed in December 1997.

When I left CERN at the end of 1998, the final phase of the LEP upgrade was still being completed, there was a vigorous ongoing LEP programme, and agreement had been obtained to operate LEP for an additional, final year in 2000. Nevertheless, it was clear that the time had come to prepare to reorganize CERN for the post-LEP era on a basis focused on the LHC project. On the face of it, things were going well, half the LHC contracts (by value) having been placed, at prices (in aggregate) just below the estimates.

It was obvious, however, that the budgetary position was extremely fragile. The deficit financing of the LHC was hyper-sensitive to small changes in the timing of contracts. Although the assumptions on timing, manpower and costs, made under great pressure in 1994 and 1996, had not been unreasonable, they had tended to be on the optimistic side, and the LHC proposal had contained no contingency funding because it would not have been accepted by some of the CERN member states. Furthermore, although the tenders for the underground civil engineering had come in below expectations, such contracts are almost always subject to revisions due to unexpected geological conditions, and the contracts for the most demanding and largest single item — the dipole magnets — had not been placed.

The LHC is an extremely challenging project, and the delays that were later produced by problems with the civil engineering and other factors should not have been a surprise. Given that it was approved in the research and development phase with no contingency, and given the 1996 discussion of risks, the 2001 revelation that the LHC would go significantly over budget should also not have been a surprise — although the lack of any warning made it a huge shock.

Lessons for the future

What lessons can be learned from the LHC saga? First, the SSC debacle strongly suggests that new projects should, if possible, be sited at existing laboratories, where they can use existing infrastructure and be spared the challenges of setting up a new laboratory and having to recruit all the key staff from scratch. Second, potential partners should be brought in at the start on equal terms or, if this is not possible, their contributions should bring added value, and they should be offered a 'voice' in the governance, as was done for the LHC. Third, approving large projects is particularly difficult for international organizations: at any time at least one partner may have economic difficulties or be out of sympathy with the organization. Fourth, stability is crucial for successful planning and execution of major projects. The events of 1996 upset orderly management, as well as shaking the confidence of CERN's staff and of potential partners in the non-member states.

Finally, it is not wise to approve projects without contingency on the basis of optimistic assumptions, although it may be worth it if — as in the case of the LHC — this is the only way to get them approved. This point was made in extreme terms, which certainly do not apply to the LHC, in a 2003 supplement to *The Times* of London on the world's great construction projects, which asserted that "If those involved didn't lie about the cost, they would never be built."

The LHC is now almost built, thanks to the dedication of the CERN staff, at a final materials cost only some 20% more than foreseen in 1993; not bad for a high-tech project approved in the research and development phase. It is a fantastic project, and I am confident that the LHC will perform superbly. ■

Chris Llewellyn Smith was Director-General of CERN from 1994 to 1998. He is currently Director of the UK Atomic Energy Authority Culham Division, Culham Science Centre, Abingdon OX14 3DB, UK.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing interests. Correspondence should be addressed to the author (chris.llewellyn-smith@ukaea.org.uk).

Building a behemoth

Oliver Brüning[1] & Paul Collier[1]

The Large Hadron Collider makes extensive use of existing CERN infrastructure but is in many respects an unprecedented undertaking. It is a proton–proton collider; therefore, it requires two separate accelerator rings with magnetic fields of opposite polarity to guide the two beams in opposite directions around its 27-km circumference. In addition, the extraordinary energies and collision rates that it has been designed to attain pose huge challenges for controlling the beam and protecting the accelerator.

The main objective of the Large Hadron Collider (LHC) is to explore the validity of the standard model of particle physics at unprecedented collision energies and rates. The design performance envisages roughly 30 million proton–proton collisions per second, spaced by intervals of 25 ns, with centre-of-mass collision energies of 14 TeV that are seven times larger than those of any previous accelerator. Reaching and maintaining this level of performance means that the LHC collider itself — although building on experiences gained at previous accelerators such as the Tevatron, at Fermilab (Batavia, Illinois), and HERA, at DESY (Hamburg) — requires a range of novel features that stretch existing technologies to the limit.

Colliders can, in principle, be designed for many different particle species (see page 270): electrons, positrons, protons, antiprotons and ions are all used in existing machines. The Tevatron, which at present defines the energy frontier for particle colliders, operates with proton and antiproton beams. By contrast, the Large Electron–Positron Collider (LEP), the last collider project at CERN, used leptons in the form of electron and positron beams. Each choice has its advantages and disadvantages. On the one hand, because leptons are elementary particles, the centre-of-mass collision energies in machines such as the LEP are precisely defined and therefore are well suited to high-precision experiments. On the other hand, the hadrons that are smashed together by the Tevatron and the LHC are composite particles, and the collisions actually occur between constituent quarks and gluons, each carrying only a proportion of the total proton energy. The centre-of-mass energy of these collisions can vary significantly, so they are not as well suited for high-precision experiments. The hadron colliders, however, offer tremendous potential for the discovery of as-yet unknown particles, because they admit the possibility of collisions over a wide range of much higher energies than is otherwise possible. Protons are relatively heavy and so lose less energy than leptons do while following a curved trajectory in a strong magnetic field. This fact, coupled with the use of superconducting magnet technology, allows the construction of a relatively compact and efficient circular machine, in which the particle beams can collide with each other at each turn. During the lifetime of the LHC, it is planned to operate with both proton and heavy-ion (lead) beams. In this review, we discuss the crucial features of the LHC that should ensure the stability and longevity of the machine while it hosts the uniquely violent collisions of these beams.

Collision energy and beam luminosity

The crucial parameters for a collider such as the LHC are the collision energy and the event rate. Taking into account the partitioning of the proton's energy between its constituent particles (that is, quarks and gluons), the choice of a proton beam energy of 7 TeV at the LHC means that average centre-of-mass collision energies will be greater than 1 TeV. To maximize the total number of events seen by the detectors, a high collision rate is also required, meaning in turn high intensities. The production rates that are achievable for antiprotons at present are too low for the design performance of the LHC; therefore, two counter-rotating proton beams are used. As a result, unlike the Tevatron, the LHC needs two separate vacuum chambers with magnetic fields of opposite polarity to deflect the counter-rotating beams in the same direction.

The number of collision events that can be delivered to the LHC experiments is given by the product of the event cross-section (which

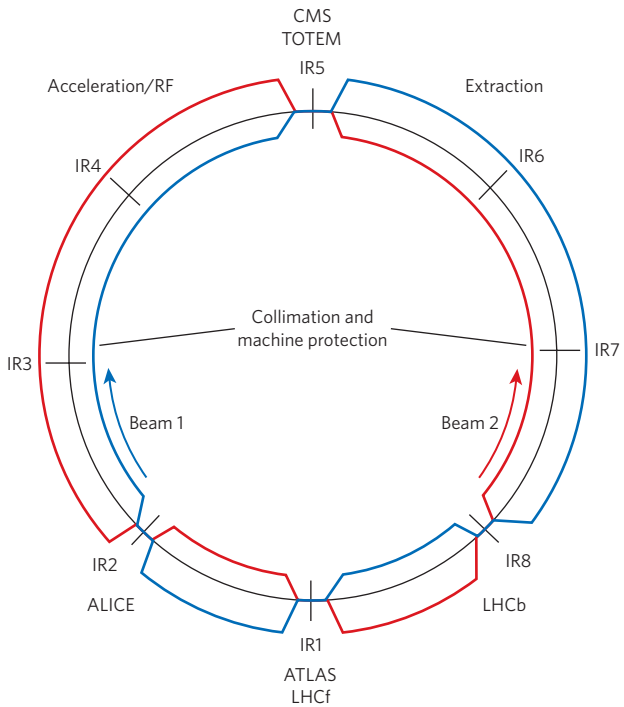


Figure 1 | Layout of the LHC collider. Two proton beams rotate in opposite directions around the ring, crossing at the designated interaction regions (IRs). Four of these (IR1, IR2, IR5 and IR8) contain the various experiments (ALICE, ATLAS, CMS, LHCb, LHCf and TOTEM). IR4 contains the radio-frequency (RF) acceleration equipment, and IR3 and IR7 contain equipment for collimation and for protecting the machine from stray beam particles. IR6 houses the beam abort system, where the LHC beam can be extracted from the machine and its energy absorbed safely.

[1]Accelerators and Beams Department, CERN, CH-1211 Geneva 23, Switzerland.

is a measure of the probability that a collision will produce a particular event of interest) and the machine luminosity, L. This is determined entirely by the proton beam parameters:

$$L = \frac{f_{rev} n_b N^2}{\sigma_x \sigma_y} F(\Phi, \sigma_{x,y}, \sigma_s)$$

Here, σ_x and σ_y are the transverse root mean squared (r.m.s.) beam sizes at the interaction points; f_{rev}, the revolution frequency; n_b, the number of particle packages ('bunches'); N, the number of particles within each bunch; and F, a geometric reduction factor that depends on the crossing angle of the two beams (Φ), the transverse r.m.s. beam size ($\sigma_{x,y}$) and the r.m.s. bunch length (σ_s). To provide more than one hadronic event per beam crossing, the design luminosity of the LHC has been set to $L = 10^{34}$ cm^{-2}s^{-1}. This translates as 2,808 bunches, each containing 1.15×10^{11} protons, a transverse r.m.s. beam size of 16 μm, an r.m.s. bunch length of 7.5 cm and a total crossing angle of 320 μrad at the interaction points. For the programme involving lead-ion collisions, L will be 10^{27} cm^{-2}s^{-1} at a centre-of-mass energy of 1,148 TeV. In this case, each ring of the LHC will contain 592 bunches, each with 7×10^7 lead ions. The transverse beam sizes will be similar to those of the proton beams.

The LEP tunnel

The LHC features six experiments (Fig. 1): two high-luminosity experiments (ATLAS[1] and CMS[2]); two supplementary experiments at low scattering angles (LHCf[3] and TOTEM[4]), which are near ATLAS and CMS, respectively; one B-meson experiment (LHCb[5]); and one dedicated ion physics experiment (ALICE[6,7]).

To make best use of the existing infrastructure at CERN, the LHC is being built in the 27-km-long LEP tunnel[8]. Approximately 22 km of the LEP tunnel consist of curved sections, or arcs, in which bending dipole magnets can be installed. The remaining 5 km consist of eight straight interaction regions that provide space for the experiments, injection and extraction elements for the proton beams, acceleration devices and dedicated 'cleaning' insertions that collimate the beam and protect the superconducting magnets from stray particles.

Dipoles and quadrupoles

Not all of the tunnel's curved sections can be used for the installation of dipole magnets. In addition to the bending fields of the dipole magnets, a circular accelerator also requires a focusing mechanism that keeps the particles centred on the design orbit. There are basically two types of circular accelerator: pulsed machines and storage rings. A storage ring is a circular accelerator where the beam may be kept for a significant time in steady conditions. In the case of the LHC, this will be several hours. Most modern storage rings use the concept of strong focusing[9,10], in which dedicated quadrupole magnets provide field components that are proportional to the deviation of the particles from the design orbit. The resulting Lorentz force prevents divergent trajectories: the particles, instead, oscillate around the design orbit as they circulate in the storage ring. The number of transverse oscillations per revolution is an important operational parameter and is referred to as the machine tune, Q. The stronger the focusing, the smaller the oscillation amplitudes (and thus the transverse r.m.s. beam size) and the larger Q is. In the longitudinal direction, the electric field supplied by a radio-frequency resonator focuses the particles into bunches and accelerates them. The LHC has two such systems, one for each beam, in one of the ring's straight sections (IR4) (Fig. 1).

The design of accelerator magnets becomes easier and less expensive for small magnet apertures, so the natural inclination is to increase the number of focusing elements in the machine to minimize the transverse beam size. But a careful balance must be struck between maximizing the

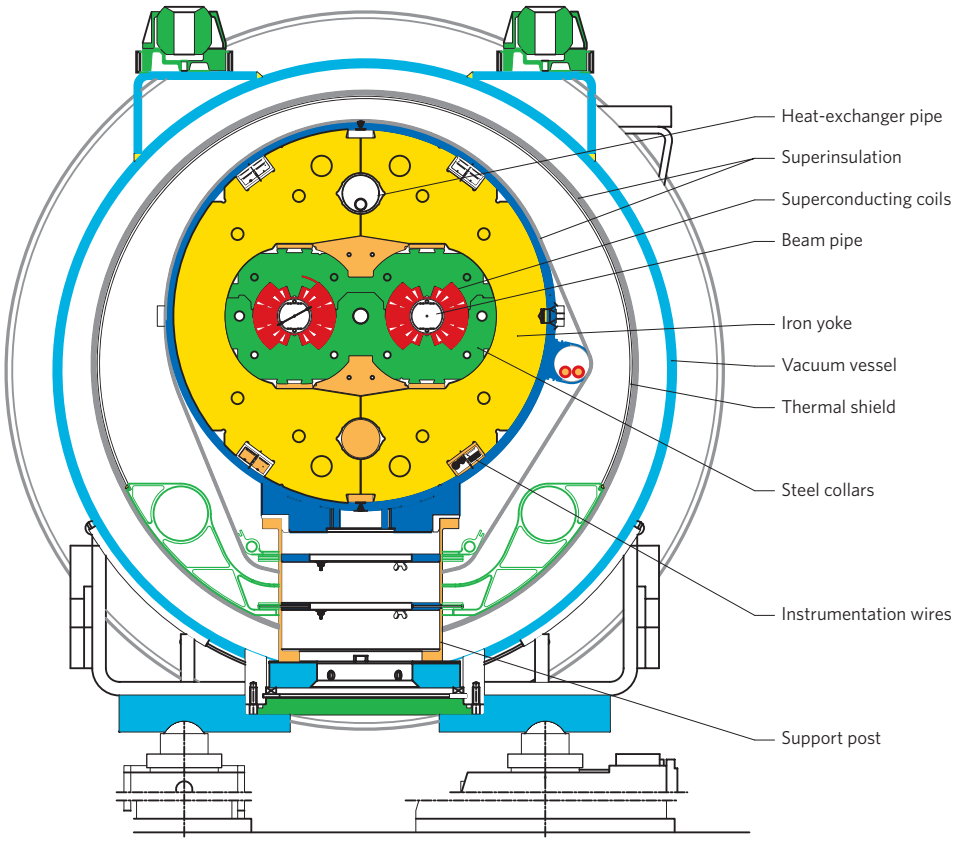


Figure 2 | Cross-section of the two-in-one design for the main LHC magnets. In the centre are the two beam pipes, separated by 194 mm. The superconducting coils (red) are held in place by collars (green) and surrounded by the magnet yoke (yellow). Together, these form the cold mass of the magnet, which is insulated in a vacuum vessel (outer blue circle) to minimize heat uptake from the surroundings.Image reproduced, with permission, from ref. 11.

space for dipole installation and providing sufficient space for transverse beam focusing. In the LHC, ~80% of the arc length is taken up by the dipole magnets, allowing the maximum transverse r.m.s. beam size in the arcs to be kept below 1.3 mm.

A two-in-one design

The combination of the length of the existing tunnel and the required beam energy sets the scale for the strength of the bending magnetic fields in the main magnets of the LHC. Keeping the 7-TeV proton beams on their closed orbits implies bending fields of 8.4 T, ~30,000 times stronger than Earth's magnetic field at its surface. Such fields are at the limit of the existing superconducting magnet technology. To confine two counter-rotating proton beams, two separate magnet apertures with opposite field orientations must be squeezed into the 3.76-m diameter of the existing LEP tunnel. The LHC therefore adopted a novel two-in-one magnet design, in which the two magnetic coils have a common infrastructure and cryostat[11] (Fig. 2).

This design provides a compact structure, with a cryostat diameter of 0.914 m, that fits two separate beam apertures into the relatively small existing machine tunnel. But it also couples the construction constraints of the two magnets, imposing new challenges and tighter tolerances on their production. This is the first time this has been done, so there is no existing experience to build on.

To minimize the number of magnet interconnections, and therefore the space lost for dipole field installations, the LHC uses 30-tonne, 15-m-long dipole magnets, which are more than twice as long as the dipole magnets in previous accelerators (~6 m for the Tevatron and HERA[12,13]). These large dimensions imposed tighter geometric constraints on the construction, transportation and installation of the magnets (Fig. 3). Each of the 8 arcs of the LHC consists of 46 repeating series of 1 quadrupole and 3 dipole magnets. Each magnet is manufactured using a niobium–titanium (NbTi)-based superconducting cable (Box 1).

Measures against magnetic quench

The operating temperature and field strength of 1.9 K and 8.4 T mean that the LHC has a very small thermal margin before the superconducting state is lost. Even small particle losses or other thermal instabilities inside the magnets can cause local heating of the material. After a section of the NbTi cable becomes a normal conductor, ohmic losses increase the operating temperature still further, an effect known as magnet quench. Testing for when a quench occurs has been an important part of the pre-installation tests of all of the LHC magnets, but efficient operation of the collider demands that the likelihood of this happening during operation is minimized.

The small tolerances for temperature fluctuations and energy deposition in the magnet coils at the LHC are combined with the extremely high energy densities inside the magnet system. The total stored electromagnetic energy — 8.5 GJ for the dipole circuits alone — is more than ten times greater than the previous record of 0.7 GJ, set by HERA[12]. The damage potential to the accelerator hardware from this stored energy is enormous: just 1 MJ is enough energy to melt 2 kg of copper.

In case of a magnet quench, this stored energy must be extracted and dissipated quickly in a controlled manner. By separating the main LHC magnet circuits into eight independent powering sectors, the stored electromagnetic energy per sector falls to that seen in existing superconducting storage rings. The drawback of this division into sectors is that it requires accurate synchronization of the different magnet sectors during operation. Existing storage rings avoid this synchronization problem by powering all main magnets in series in a central circuit. The LHC will enter new territory in this respect.

Damage to individual magnet units during a quench is avoided by a dedicated magnet protection system that monitors the voltage drop across each magnet unit. As soon as any part of the magnet cable loses its superconducting state, the voltage drop across the magnet will become non-zero. This jump will activate special heaters inside the magnet to bring the whole magnet into a normal conducting state, thus spreading the quench over the whole magnet length. A dedicated quench diode dissipates the stored electromagnetic energy before it can damage the magnet coils.

Figure 3 | Installing the LHC magnets. a, An LHC dipole ready for installation at the CERN site. **b**, Transport of LHC magnets in the tunnel, alongside installed elements, illustrating the tight space conditions for installation. Images reproduced with permission from CERN.

The stored energy in the proton beams themselves is another dangerous source of energy deposition in the superconducting magnet coils. At 7 TeV and an intensity of 3.23×10^{14} protons, the kinetic energy of each of the LHC beams is 362 MJ. Safe beam extraction in case of problems during machine operation, or at the end of a period of operation for data taking by the experimental detectors (physics fill), is assured by two installations: the beam abort system, and the machine protection system. The ring of the LHC has a dedicated beam abort system, formed of specially designed absorber blocks capable of absorbing the full beam intensities at 7 TeV without damage. The machine protection system constantly monitors all critical machine parameters and initiates a beam abort if the parameters exceed the acceptable operation tolerances or if the beam losses along the storage ring become too large.

Beam lifetimes

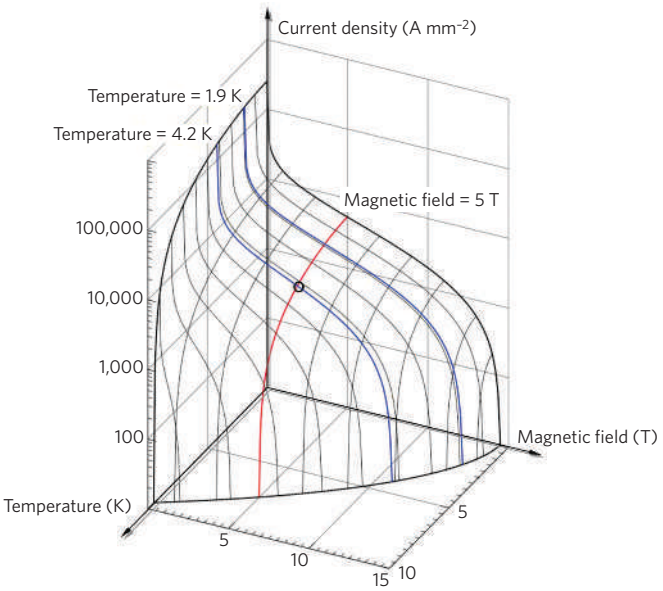
Beam intensity — and hence luminosity — decays during the operation of an accelerator in colliding mode. After these parameters become too small for efficient operation, the beams are discarded using the beam abort system, and a new fill of proton beams needs to be prepared, injected and accelerated. One of the main and unavoidable causes of reductions in beam intensity is the collisions inside the detectors themselves, because these cause the disintegration of beam particles. The rate of this disintegration is given by the product of the machine luminosity, the total cross-section for an inelastic interaction of two protons at 7 TeV, and the number of collision points. Assuming a total inelastic cross-section of 10^{-25} cm^2 at 7 TeV and two main interaction points with

Box 1 | The LHC superconductor

The superconducting magnets of the LHC are superlative devices: had the LHC been made of conventional magnets, it would have needed to be 120 km long to achieve the same energies, at the cost of a much greater electricity consumption.

Like all superconductors, NbTi (the material used for the cables of the LHC magnets) is a superconductor only if its operational parameters — temperature, current density and ambient magnetic field — are within certain bounds[15–17]. The critical surface below which this combination of parameters must lie is shown in the figure. At the preferred operating temperature for most existing superconducting accelerators such as HERA and the Tevatron[11,12], 4.2 K, the critical magnetic field is around 5 T.

The temperature of the LHC, 1.9 K, allows the generation of the required 8.4-T magnetic field using a current density of 1.5–2 kA mm^{-2} inside the superconducting cables. It also allows the use of superfluid helium, which has high thermal conductivity, as a coolant. A helium inventory of 120 tonnes or more will be needed to cool the total magnet mass of 37,000 tonnes, the largest such inventory in the world. Figure courtesy of L. Bottura (CERN).



a luminosity of 10^{34} $cm^{-2}s^{-1}$, the beam intensities will have dropped to half of their initial values after ~45 hours.

Particles are also lost through perturbations and resonances in the proton motion that deflect particles away from the design orbit. These are generated, for example, at the collision points, where a particle in one beam is exposed to the Coulomb field of the opposing beam, or by field imperfections in the main magnets. Thanks to the focusing mechanism of the quadrupole magnets, these deflections do not lead directly to particle losses but, initially, just to an oscillation around the design orbit. But consecutive perturbations can add up coherently if the particle oscillations are in resonance with the revolution frequency, in which case the oscillation amplitudes can grow until the particles are lost when they reach the boundary of the LHC vacuum system.

Two approaches minimize this amplitude growth. First, extreme care is taken during the magnet design, construction and installation in order to minimize any imperfections in the machine. Second, the LHC is equipped with dedicated circuits that allow the correction of the most dominant residual field errors. All magnets are measured before their installation to develop an accurate magnetic model of the entire machine's operation. In total, the LHC features 112 correction circuits per beam (not including simple steering magnets for an adjustment of the central orbit), and all of these must be adjusted during operation. To compound the difficulty, the field errors of a superconducting magnet are not constant but vary with time as a function of the magnet's powering history.

Other causes of a reduction in luminosity during operation include the scattering of protons on residual gas molecules inside the beam vacuum system, and the Coulomb scattering of the protons inside each bunch as they perform longitudinal and transverse oscillations while circulating inside the storage ring. The rate of collisions with residual gas molecules depends on the pressure and gas composition inside the machine vacuum system. An efficient operation with proton beams requires vacuum levels below 10^{15} molecules per cubic metre for all gas components (H_2, He, CO, CO_2 and so on), corresponding to a pressure of less than 10^{-7} Pa at 5 K. (Atmospheric pressure is ~10^5 Pa at sea level.) This, in turn, demands an elaborate system of different vacuum pumps. In its final phase, the pumping mainly relies on the cryo-pumping of the cold surfaces that exist at the boundary between the beam vacuum and the helium in the superconducting magnets, similar to the way that ice builds up on the surfaces of the freezing compartment of a household refrigerator.

Collimation

There are therefore several unavoidable mechanisms causing a continuous loss of particles during LHC operation through a relatively slow drift to larger oscillation amplitudes. Two dedicated collimation insertions with specially designed absorber blocks mop up these stray particles before they can reach the cold aperture of the superconducting magnets (and so possibly cause a magnet quench). This mopping up must be done with high efficiency so that only 1 in 10,000 particles that hit the primary collimators end up inside the cold aperture. Such a high cleaning efficiency requires extremely tight tolerances for the main machine parameters during operation, as well as the use of a complex two-stage collimation system with additional dedicated absorbers at crucial locations. The LHC is the first high-energy collider that requires a beam collimation during all stages of the operation to protect its machine elements — previous colliders only required a beam collimation during the physics run, mainly to reduce the background in the experiments. For additional safety, therefore, the collimator jaws at the LHC are made of fibre-reinforced graphite so as to be able to withstand the direct impact of a large proportion of the 7 TeV beam.

Working up to full beam strength

After the LHC proton beams have been prepared and injected into the accelerator using the existing accelerators at CERN[14] (Box 2), the acceleration can be initiated. This acceleration relies on a synchronous change of the machine settings with the increasing dipole field. In the case of the LHC, the final beam energy is more than 15 times greater than that of the injected beam (7,000 GeV compared with 450 GeV). With a high impedance in the main magnet circuits, the process of increasing the magnet current, and therefore the energy, is slow in the LHC, taking ~20 minutes. During this operational phase, the transverse beam dimensions shrink as the rigidity of the beam increases.

The injection and acceleration takes place with the beams separated in the experimental regions and with a lower focusing strength in these areas than in the final configuration for luminosity production. After reaching high energy, a synchronized change in the settings of the focusing elements is made at each interaction point to reduce the beam spot size at the interaction point. As a result, the beam size in the adjacent final focusing quadrupoles increases. In the final configuration, the aperture of these elements is smaller than in the rest of the machine and must be protected by further reducing the collimation gap.

The final step before the experiments can begin taking data is to remove the separation scheme and to bring the beams into collision in each experimental area. Careful optimization is required to align the beams correctly and to maximize the overlap of the 16-μm beam spots.

Once data taking has started, the luminosity will decrease as the intensity falls. In fact, because the luminosity is proportional to the square of the intensity, a reduction in the intensity by ~30% will halve the luminosity. The goal for efficient machine operation is a luminosity lifetime that is considerably longer than the average time for preparing a new fill of proton beams. Assuming an exponential decay of luminosity, and an

intensity lifetime of 45 hours, the luminosity lifetime will be ~15 hours for the LHC. The overall collider efficiency depends on the ratio of the run length and the average turnaround time. Assuming a 5-hour turnaround time, the optimum run length will be ~10 hours.

The commissioning process

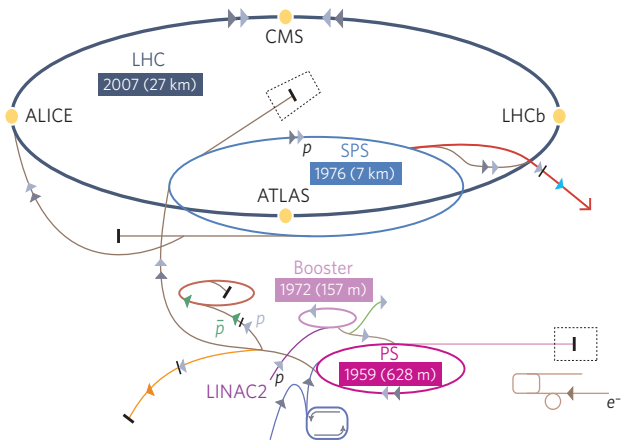
The LHC is a huge, complex facility, and careful and precise control of all machine elements is necessary. Commissioning the whole machine is a challenge in itself. Careful commissioning of each individual set of accelerator hardware will be followed by rigorous system tests and integrated operation of the whole accelerator before any beam is injected. In the early phases of beam operation, the complexity can be reduced by limiting the number of bunches, the intensity per bunch and even the final energy. At each stage in the commissioning process, the equipment and protection systems must be tested and run to allow operation with beam while minimizing the risk of damage to the accelerator itself.

For the first operation of the LHC at 7 TeV, there will be a single bunch in each ring. From there, a staged increase in the number of bunches is intended, with schemes for 43, 156 and 936 bunches per ring envisaged before arriving at the final number of 2,808 bunches per ring. The simplest scheme, with 43 bunches per ring and an intensity per bunch around half the nominal value, represents a stored energy that is already comparable to that of the Tevatron.

Box 2 | Preparing the LHC beam

The beam of the LHC starts off in a 50-MeV linear accelerator, LINAC2 (see figure). It is then passed to a multi-ring booster synchrotron for acceleration to 1.4 GeV, and then to the 628-m-circumference Proton Synchroton (PS) machine to reach 26 GeV. During acceleration in the PS, the bunch pattern and spacing needed for the LHC are generated by splitting the low-energy bunches. A final transfer is made to the 7-km Super Proton Synchroton (SPS) machine, where the beam is further accelerated to 450 GeV. At this point, it is ready for injection into the LHC. The cycle takes ~20 s and creates a ribbon, or train of bunches, with a total kinetic energy of more than 2 MJ. This is ~8% of the beam needed to fill an LHC ring completely, so the whole cycle must be repeated 12 times per ring.

The transfer of the bunch trains from the SPS to the LHC is one of the most dangerous phases of the operational cycle of the LHC. The injected beam already has sufficient energy to damage the LHC equipment, and the transfer involves the use of fast kicker magnets to abruptly change the trajectory of the beam to move it out of the SPS, down a 3-km transfer line, and into the LHC. Any mis-steering here could be disastrous, so a low-intensity 'pilot' beam is injected into the machine first. This is used to measure and correct the machine parameters before the full-intensity injection sequence is allowed to start. Each injection is positioned in the LHC circumference so as to generate the complete pattern for each beam. During the 8 minutes needed to fill the LHC completely, the stability of the whole complex is critical and must be carefully monitored. Figure modified with permission from CERN.



During the first full year of LHC operation, the number of bunches and the intensity per bunch will be increased slowly. It is hoped that a luminosity of 10^{33} cm^{-2}s^{-1}, or 10% of the nominal value, will be reached during this time. In subsequent runs, the performance will be slowly increased towards the nominal value as understanding of the machine and control of the machine parameters is refined.

The LHC is a machine in which all technologies are stretched towards their limit, and it has been built, in many cases, with very small operational margins in the equipment. It is probable that upgrades to certain accelerator components will be made during the lifetime of the machine. Some of these will be designed to re-introduce operational margins in crucial areas in which machine efficiency can be improved. Others will be designed to increase the nominal performance of the machine.

The outlook

The LHC is designed to push back the frontiers of our knowledge of fundamental particle physics. With the requirement of providing both high energies and high beam intensities, there are many challenges that had to be overcome to produce a viable design for the complete machine. Realizing the designs for each component of the accelerator has often, in turn, pushed back the technical boundaries for the design and performance of the individual accelerator systems. The sheer size and complexity of the complete machine makes the commissioning and operation of the LHC a challenge in itself. But its many technical innovations mean that the LHC should be capable of helping us to explore — and, we hope, answer — some of the most fundamental questions in particle physics today. ■

1. Armstrong, W. W. *et al.* (The ATLAS Collaboration). ATLAS: *Technical Proposal for a General-Purpose* pp *Experiment at the Large Hadron Collider at CERN. CERN* <http://cdswebdev.cern.ch/record/290968> (1994).
2. Della Negra, M., Petrilli, A., Hervé, A. & Foà, L. *CMS Physics: Technical Design Report. Vol. 1: Detector Performance and Software. CERN* <http://cdswebdev.cern.ch/record/922757> (2006).
3. Mukari, Y., Itow, Y. & Sako, T. *LHCf Experiment: Technical Design Report. CERN* <http://cdswebdev.cern.ch/record/926196> (2006).
4. Berardi, V. *et al. Total Cross-Section, Elastic Scattering and Diffractive Dissociation at the Large Hadron Collider at CERN: TOTEM Technical Design Report. CERN* <http://cdswebdev.cern.ch/record/704349> (2004).
5. Amato, S. *et al.* (The LHCb Collaboration). *LHCb Technical Proposal, a Large Hadron Collider Beauty Experiment for Precision Measurements of CP Violation and Rare Decays CERN/LHCC 98-4. CERN* <http://cdsweb.cern.ch/record/622031> (1998).
6. Carminati, F. *et al.* (The ALICE collaboration). ALICE: physics performance report, volume I. *J. Phys. G* **30,** 1517–1763, doi:10.1088/0954-3899/30/11/00 (2004).
7. Alessandro, B. *et al.* (The ALICE collaboration). ALICE: physics performance report, volume II. *J. Phys. G* **32,** 1295–2040, doi:10.1088/0954-3899/32/10/001 (2006).
8. CERN. *LEP Design Report. CERN* <http://cdsweb.cern.ch/record/102083> (1984).
9. Christofilos, N. C. Focusing system for ions and electrons. US patent 2,736,799 (1950); reprinted in Livingston, M. S. *The Development of High Energy Accelerators* (Dover, New York, 1966).
10. Courant, E. D. & Snyder, H. S. Theory of the alternating gradient synchrotron. *Ann. Phys. (Leipz.)* **3,** 1–48 (1958).
11. Brüning, O. S. *et al.* (eds) *LHC Design Report Vol. 1: The LHC Main Ring. CERN* <http://cdsweb.cern.ch/record/782076> (2004).
12. Cole, F. T. *et al. A Fermilab Superconducting Accelerator Design Report*. Beams Document 1888, Fermilab, Batavia Illinois <http://beamdocs.fnal.gov/AD-public/DocDB/ShowDocument?docid=1888> (1979).
13. *HERA — A Proposal for a Large Electron Proton Colliding Beam Facility at DESY.* DESY Hamburg, Germany. *SLAC* <http://www.slac.stanford.edu/spires/find/hep/www?key=897230> (1981).
14. Benedikt, M., Collier, P., Mertens, V., Poole, J. & Schindl, K. (eds) *LHC Design Report Vol. 3: The LHC Injector Chain. CERN* <http://cdsweb.cern.ch/record/823808> (2004).
15. Wilson, M. N. *Superconducting Magnets* (Clarendon, Oxford, 1983).
16. Bottura, L. A practical fit for the critical surface of NbTi. *IEEE Trans. Appl. Supercon.* **10,** 1054–1057 (2000).
17. Leroy, D. Review of the R&D and supply of the LHC superconducting cables. *IEEE Trans. Appl. Supercon.* **16,** 1152–1159 (2006).

Acknowledgements

The authors would like to acknowledge the many thousands of people at CERN and in collaborating institutes who have contributed to the design and construction of the LHC machine and experiments.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to P.C. (paul.collier@cern.ch).

NATURE|Vol 448|19 July 2007|doi:10.1038/nature06078

Detector challenges at the LHC

Steinar Stapnes[1,2]

The best way to study the existence of the Higgs boson, supersymmetry and grand unified theories, and perhaps the physics of dark matter and dark energy, is at the TeV scale. This is the energy scale that will be explored at the Large Hadron Collider. This machine will generate the energy and rate of collisions that might provide evidence of new fundamental physics. It also brings with it the formidable challenge of building detectors that can record a large variety of detailed measurements in the inhospitable environment close to the collisions points of the machine.

Four main experiments have been designed and constructed for the Large Hadron Collider (LHC) machine: ATLAS, CMS, LHCb and ALICE. ATLAS and CMS are large general-purpose experiments. LHCb will study *b*-quark systems, produced predominantly in the forward direction, and ALICE is designed specifically for studies of heavy-ion collisions (see page 302).

This review focuses on the challenges — related to tracking, calorimetry, muon detection, triggering and data acquisition — faced by the designers and builders of the general-purpose detectors ATLAS and CMS, as well as some of the particular issues for the more specialized detector LHCb.

Experimental measurements at the LHC

Inside the 27-km ring of the LHC, bunches of 10^{11} protons will collide 40 million times per second to provide 14-TeV proton–proton collisions at the LHC design luminosity of 10^{34} cm^{-2}s^{-1} (see page 285). With an inelastic proton–proton cross-section of about 100 mb (~10^{-25} cm^2), this gives 25 events per bunch crossing, or a total rate of 10^9 inelastic events per second. This means that around 1,000 charged particles will emerge from the collision points every 25 ns, within a volume defined by $|\eta| < 2.5$, where pseudorapidity, η, is related to the polar angle relative to the beam axis, θ, by $\eta = -\ln[\tan(\theta/2)]$ (Fig. 1a).

This formidable luminosity and interaction rate are necessary, because the expected cross-sections are small for many of the LHC benchmark processes (such as Higgs production and decay, and some of the processes needed to search for and explore new physics scenarios such as supersymmetry and extra dimensions). They also raise a serious experimental difficulty; every candidate event for new physics will, on average, be accompanied by 25 inelastic events occurring simultaneously in the detector.

The very nature of proton–proton collisions creates a further difficulty. The cross-sections for producing jets of particles, through quark interactions governed by quantum chromodynamics (QCD), are large compared with the rare processes being sought — several orders of magnitude larger, even for jet production above 500 GeV. Therefore, one has to look for characteristic experimental signatures, such as final states involving one or more leptons, or photons, or with missing transverse energy or secondary vertices, to avoid being drowned by QCD background processes. Searching for such final states among already rare events imposes further demands on the luminosity needed and on the detectors' particle identification capabilities.

Specific requirements for the LHC detector systems[1–8] have been defined using a set of benchmark processes that covers most of the new phenomena that one might hope to observe at the TeV scale. The first such process is the production of the standard-model Higgs boson, which is particularly important because there is a wide range of decay modes possible, depending on the mass, m_H, of the Higgs boson. If m_H is low (less than 180 GeV, which is twice the mass of the *Z* boson), the natural width is only a few MeV, and the observed width will be defined by the instrumental resolution. The dominant decay mode into hadrons is difficult to isolate, because of the QCD background. Therefore, the two-photon decay channel will be important, as will other channels, including associated productions such as *ttH*, *WH*, *ZH* (see page 270), for which a lepton from the decay of the accompanying particle will be used for triggering and background rejection.

Above 130 GeV, Higgs decay into *ZZ* (one Z being virtual when m_H is below the *ZZ* threshold), with its four-lepton final state, will be the most interesting channel. Above 600 GeV or so, *WW* or *ZZ* decays into jets or into states involving neutrinos (leading to missing transverse energy because the neutrinos are undetected) are needed to extract a signal; for m_H close to 1 TeV, it becomes necessary to tag 'forward' jets, in the region $2 > |\eta| > 5$, from the *WW* or *ZZ* fusion production mechanism. The Higgs might not even be of the standard-model variety; detection of some of the Higgs particles of the 'minimal supersymmetric extension of the standard model' (MSSM) would require very good sensitivity to processes involving tau leptons and *b* quarks.

If supersymmetric particles such as squarks and gluinos are produced at the LHC, their decays would involve cascades that always contain a lightest stable supersymmetric particle, or LSP (if R-parity is conserved). Because the LSP interacts very weakly with the detector, the experiments would measure a significant missing transverse energy in the final state. The rest of the cascade results in a number of leptons and jets.

Several new models, motivated by theories of quantum gravity, propose the existence of extra dimensions[1–4]. In terms of experimental signatures, the emission of gravitons that escape into extra dimensions would result in missing transverse energy; furthermore, Regge-like excitations could manifest themselves as *Z*-like resonances with ~TeV separations in mass. Other experimental signatures could be anomalous, high-mass dijet production and mini-black-hole production with very spectacular decays involving democratic production of jets, leptons, photons, neutrinos, and *W* and *Z* bosons.

The LHC will also allow studies of QCD, electroweak and flavour physics. For example, *t* quarks will be produced at the LHC at a rate measurable in hertz. New, heavy gauge bosons (W' and Z') could be accessible at masses up to 5–6 TeV. To study their leptonic decays, high-resolution lepton measurements and charge identification are needed

[1]Department of Physics, University of Oslo, 0316 Blindern, Oslo, Norway. [2]Department of Physics, CERN, CH-1211 Geneva, Switzerland.

up to transverse momenta of a few TeV. Another new-physics signature could be jets produced with very high transverse momenta; if quarks are composite rather than fundamental particles, deviations from QCD expectations in the jet cross-sections could result.

Detector concepts

The necessary detection capabilities for all of these experimental signatures lead to a stringent set of design requirements. First of all, owing to the experimental conditions at the LHC, the detectors need fast, radiation-tolerant electronics and sensor elements. In addition, high granularity of the detectors is needed to be able to handle the particle fluxes and to reduce the influence of overlapping events. Good charged-particle momentum resolution and reconstruction efficiency in the tracking system, and in the inner tracker specifically, are essential. For efficient high-level triggering and offline tagging of taus and *b* quarks (which decay a short distance from the primary interaction vertex at which they are produced), pixel detectors close to the interaction region are needed to observe the distinctive secondary vertices.

Two key requirements are good electromagnetic calorimetry for electron and photon identification and measurements, and full-coverage hadronic calorimetry for accurate jet and missing-transverse-energy measurements. Likewise, good muon identification and momentum resolution over a wide range of momenta, and the ability to determine unambiguously the charge of muons with high transverse momentum, are essential. Finally, triggering the event readout on the presence of leptons, jets, photons or missing transverse energy — and at low transverse-momentum thresholds to ensure high efficiencies for most of the physics processes of interest at LHC — is an absolute requirement to reduce the data rate (a few hundred collisions out of the 40 million taking place every second are finally kept) to a level that can be handled offline.

The layout of the ATLAS detector[1,2] is shown in Fig. 1c. It has an inner, thin, superconducting solenoid surrounding the inner detector cavity, and large, superconducting, air-core toroids, consisting of independent coils arranged with an eight-fold symmetry, outside the calorimeters. The inner detector comprises a large silicon system (pixels and strips) and a gas-based transition-radiation 'straw' tracker. The calorimeters use liquid-argon technology for the electromagnetic measurements and also for hadronic measurements in the endcaps of the detector. An iron/scintillator system provides hadronic calorimetery in the central part of the detector. The muon system is based on gas detectors and has precise tracking chambers and trigger chambers for a robust and efficient muon trigger. The ATLAS detector has a radius of 13 m and is 46 m long, with a weight of 7,000 tonnes.

The design of CMS[3,4] is shown in Fig. 1b . The main distinguishing features of CMS are a high-field solenoid housing a full silicon-based inner tracking system (pixels and strips), a fully active, scintillating crystal electromagnetic calorimeter, and a compact scintillator/brass hadronic calorimeter. Outside the solenoid, there is a hadronic 'tail-catcher' in the central region, and an iron-core muon spectrometer sitting in the return field of the powerful solenoid, with tracking chambers and trigger

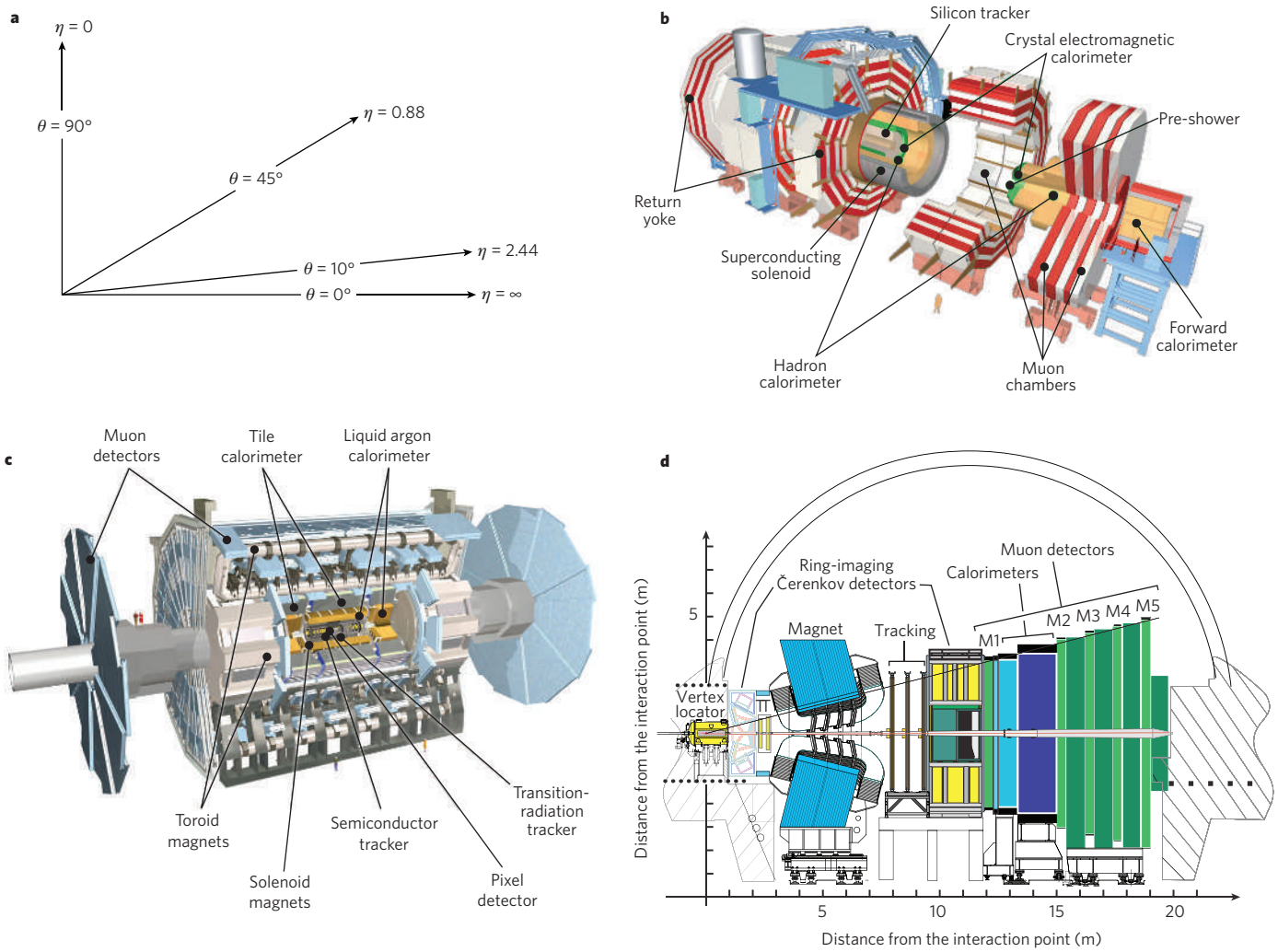


Figure 1 | Detector design. The complex experimental apparatus comprises different detector components, each optimized for a particular task. Regions of the detector volume are commonly described using the variable pseudorapidity, η, which is related to the polar angle, θ, as shown in **a**. The geometry and basic elements of the general-purpose LHC detectors, CMS (**b**) and ATLAS (**c**), are similar, but the layout of the more specialized detector LHCb (**d**) is optimized for detecting the production of *b* quarks in the forward direction. Images b–d reproduced with permission from CERN.

chambers. The CMS system is more compact than ATLAS, and has a radius of 7.5 m and length of 24 m, but weighs 12,000 tonnes.

LHCb[8] is designed to operate at a luminosity of 2×10^{32} cm^{-2}s^{-1}, and its instrumentation is concentrated in the forward direction, between 10 and 300 mrad (Fig. 1d), because this is the region in which the pairs of b and $\bar{b}$ quarks that it aims to study are predominantly produced. The LHCb detector has a silicon vertex detector around the interaction region; then a tracking system consisting of silicon microstrip detectors and a straw tracker, and it includes a dipole magnet. It also has two ring-imaging Čerenkov detectors, positioned in front of and after the tracking system, for charged-hadron identification; a calorimeter system and finally a muon system. LHCb, in particular, has to trigger efficiently on the secondary vertices that are the signature of b quarks, and so its vertex and tracking detectors are factored into an early stage of its trigger scheme.

To construct these large detectors requires substantial resources. The ATLAS and CMS communities each consist of more than 150 universities and institutions from about 35 countries with about 2,000 collaborators per experiment. (LHCb is a factor of three smaller.) Research and development for the LHC detectors began around 1990; the construction projects were approved in 1996 and started in earnest around 1998. The effort to build all of the detector components has involved physicists all over the world, with groups of geographically distributed institutes taking responsibility for the construction of various parts according to their specific expertise and capabilities, as well as involving a large network of industrial partners.

Inner detectors

The ATLAS and CMS inner detectors (Fig. 2) are contained in central solenoid fields of 2 T and 4 T, respectively. They provide efficient tracking of charged particles within the pseudorapidity range $|\eta| < 2.5$, allowing momentum measurement and the reconstruction of primary and secondary vertices. Both systems are largely based on silicon, with high-granularity pixel systems at the smallest radii, and silicon-strip detectors at larger ones. ATLAS has a 'straw' tracker at the largest radius.

Silicon detectors are p–n junction diodes that are operated at reverse bias[9]. This forms a sensitive region depleted of mobile charge and sets up an electric field that sweeps charge (electron–hole pairs) liberated by radiation towards the electrodes. Detectors typically use an asymmetric structure, for example, a highly doped p electrode and a lightly doped n region (p-i-n), so that the depletion region extends predominantly into the lightly doped volume. By adding highly doped n electrodes (n-i-n) at the back, the back side can also be read out. Integrated circuit technology allows the formation of high-density micrometre-scale electrodes on large (10–15 cm in diameter) wafers, providing excellent position resolution. Furthermore, the density of silicon and its small ionization energy (the energy needed to created an electron–hole pair) result in adequate signals with active layers only 200–300 μm thick, and the charge mobility is such that the signals are also fast (typically tens of nanoseconds).

The main challenges for the inner detector parts are the high particle rates, the radiation tolerance needed and the control of ageing effects. The ATLAS and CMS trackers had to be designed to withstand high radiation doses (500–1,000 kGy) for the innermost pixel layers, and up to 100 kGy for the systems farther away from the interaction point, after 10 years of operation). As a result, the development of the integrated front-end electronics for these systems has been a major problem to solve over several years and design iterations. These circuits must be fast, radiation tolerant and low power, and are integrated on low-mass modules where cooling and material limitations are severe. Several rounds of testbeam measurements and rigorous irradiation programmes have been necessary to prove that the circuits will function in their final assemblies, as well as after high irradiation.

A similarly stringent research and development programme was needed for the silicon sensors themselves[10,11], for which the major difficulty is bulk radiation damage. The relevant parameter is the accumulated dose in the volume of the inner detector, which varies between 10^{15} and 10^{14} cm^{-2} from the innermost layers to the outer ones (where n_{eq} means the number of equivalent particles, normalized using non-ionizing energy loss cross-sections to the damage expected to be caused by 1 MeV neutrons). These radiation doses have severe consequences for the silicon sensors (as they do for all other module components and for the thermal design of the system). They cause increased leakage current and changes in effective doping, and therefore changes in depletion and operation voltages, leading to type inversion for n-type sensors. After type inversion, the effective doping also shows an increase with time following irradiation (reverse annealing) that is temperature dependent. To maintain the operation voltage within reasonable limits, the sensors are therefore kept cold (–10 °C to 0 °C) throughout their lifetime, which has the added benefit of reducing the leakage current.

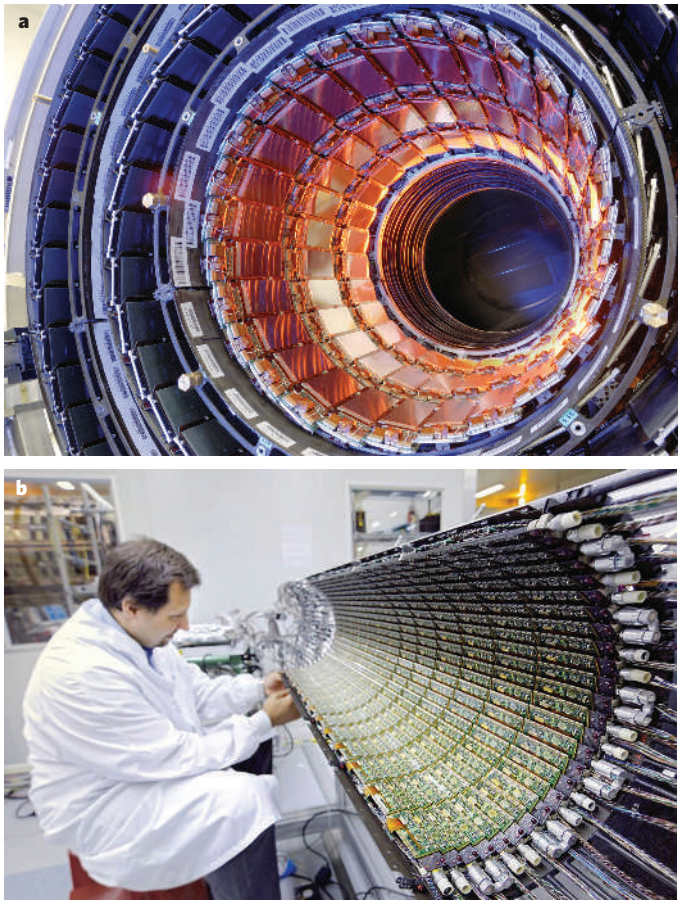


Figure 2 | Tracking systems. a, Silicon microstrip detectors in the CMS barrel region. **b**, The ATLAS pixel detector during the final assembly stage. Images reproduced with permission from CERN.

All of these effects have been carefully mapped out, and various design options have been evaluated in prototypes. Of particular interest are n-i-n silicon sensors, in which the charge-collection region grows with bias voltage from the n-implant side after type inversion following irradiation. Therefore, high efficiency can be obtained from an under-depleted detector. This allows a system to be specified with a lower maximum operating voltage. The ATLAS and CMS pixel systems and the LHCb vertex detector use such sensors. These require double-sided processing and are relatively complex and costly, so for the large-area silicon-strip systems, simpler, single-sided p-i-n designs have been adopted.

Considering the flux of charged particles at increasing radii around the LHC beams, three detector regions are defined in ATLAS and CMS. In the first of these, closest to the interaction point where the particle flux is highest, there are silicon pixel detectors (Fig. 2b), whose cell sizes of 50×400 μm^2 and 100×150 μm^2 in ATLAS and CMS, respectively, give an occupancy of about 10^{-4} per pixel per bunch crossing. To improve the measurement of secondary vertices (typically from b-quark decays) an innermost layer of pixels has been introduced as close to the beam as is practical, at a radius of about 4.5 cm. The lifetime of this layer will

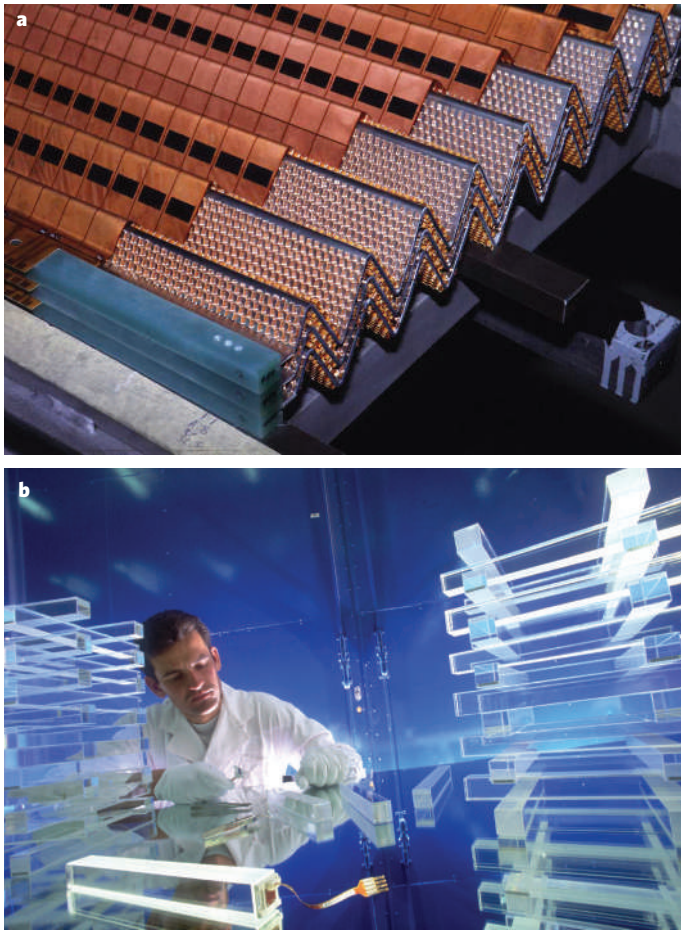


Figure 3 | Calorimetry: different approaches. a, The layers of the ATLAS electromagnetic calorimeter have an 'accordion' geometry. **b**, Tens of thousands of lead tungstate crystals have been prepared and tested, before being assembled into the electromagnetic calorimeter of the CMS detector. Images reproduced with permission from CERN.

be limited, owing to radiation damage, and it may need to be replaced after a few years. The pixel systems in ATLAS and CMS are very much larger than any comparable existing system. The ATLAS pixel system covers about 2 m^2 and has 80 million channels; the CMS pixel system is only slightly smaller.

Pixel detectors are expensive and have high power density, so at a certain radius and system size, silicon microstrip systems become the preferred technology. In the intermediate tracking region of ATLAS and CMS, at a radius of 20–55 cm, the particle flux becomes low enough to use silicon microstrip detectors. Barrel cylinders and endcap discs provide coverage out to about $|\eta| = 2.5$. Strip dimensions of 10–12 cm × 80–120 μm lead to an occupancy of 1–3% per bunch crossing. Both trackers use stereo angle in some of the strip layers (that is, strips placed at a small angle with respect to the z axis, 40 mrad and 100 mrad for ATLAS and CMS, respectively) to improve the resolution in z. In these microstrip systems, it has been essential to find a good balance between the pitch of the cells (determining resolution and occupancy), radiation effects, capacitive load (noise), material length and costs.

Finally, in the outermost region (beyond about 55 cm), the particle flux has dropped sufficiently to allow the use of larger-pitch silicon microstrips in the CMS tracker, with a maximum cell size of 25 cm × 180 μm while keeping the occupancy to about 1%. There are six layers of such microstrip modules in the barrel, accompanied by nine endcap discs providing coverage out to about $|\eta| = 2.5$, amounting to 15,400 modules and 9.6 million channels, and spanning a total detector area of more than 200 m^2.

For ATLAS, at radii greater than 56 cm, a large number of tracking points (typically 36 per track) is provided by the 'straw' tracker — 300,000 straw-tubes embedded in fibre or foil radiators and filled with a xenon-based gas mixture. This detector allows continuous track following with less material per point, and also has electron identification capabilities. X-ray photons are produced through transition radiation because highly relativistic particles such as electrons traverse the straw tracker's multiple interfaces.

Another silicon microstrip detector at LHC, the Vertex Locator[12] of LHCb, has some special features that present significant challenges. The 42 double-sided, half-moon-shaped detector modules are placed at a radial distance from the beam (8 mm) that is smaller than the aperture required by the LHC during injection and must therefore be retractable. For minimizing the material between the interaction region and the detectors, the silicon sensors are inside a thin aluminum box at a pressure of less than 10^{-4} mbar (10^{-2} Pa). The n-i-n sensors used have rφ geometry with pitch (40–140 μm) depending on the radius.

Calorimeters

The calorimeters absorb and measure the energies of electrons, photons and hadrons. In the design of the electromagnetic calorimeters for both ATLAS and CMS, the emphasis is on good resolution for photon and electron energy, position and direction measurements, and wide geometric coverage (up to $|\eta|$ close to 3.0). In the QCD-dominated environment of the LHC, the ability to reject neutral pions is crucial for photon and electron identification. It is also important to have efficient photon and lepton isolation measurements at high luminosities. For the hadronic calorimeters, the emphasis is on good jet-energy measurements, and full coverage (to $|\eta| = 5$) to be able to ascribe the observation of significant missing transverse energy to non-interacting particles (such as neutrinos, or light neutralinos from supersymmetric-particle cascade decays) rather than to losses in the forward regions. Last but not least, the quantities measured in the calorimeters play a crucial part in the trigger of the experiment as signatures of significant parts of the new physics sought at the LHC.

These considerations bring stringent requirements for high granularity and low noise in the calorimeters. The major technical difficulties for the calorimeters are related to the radiation doses (reaching 200 kGy for the electromagnetic part and 1,000 kGy for the hadronic part at the highest η), the sampling speed and the dynamic range needed to measure with low noise and good resolution over a wide energy range.

The ATLAS calorimetry consists of an electromagnetic calorimeter covering the pseudorapidity region $|\eta| < 3.2$, a hadronic barrel calorimeter covering $|\eta| < 1.7$, hadronic endcap calorimeters covering $1.5 < |\eta| < 3.2$, and forward calorimeters covering $3.1 < |\eta| < 4.9$, as shown in Fig. 1c. Over the pseudorapidity range $|\eta| < 1.8$, a presampler is installed in front of the electromagnetic calorimeter to correct for energy loss upstream.

The electromagnetic calorimeter system consists of layers of lead (creating an electromagnetic shower and absorbing particles' energy), interleaved with liquid argon (providing a sampling measurement of the energy-deposition) at a temperature of 89 K. The system's 'accordion' geometry[1,2] provides complete azimuthal symmetry, without cracks, and has been optimized for the high sampling rate environment of the LHC (Fig. 3a). The barrel section is sealed within a barrel cryostat, which also contains the central solenoid, surrounding the inner detector. The endcap modules are contained in two endcap cryostats that also contain the endcap hadronic and forward calorimeters.

The hadronic barrel calorimeter is a cylinder divided into three sections: the central barrel and two identical extended barrels. It is again based on a sampling technique, but uses plastic scintillator tiles embedded in an iron absorber. The vertical tile geometry makes it easier to transfer the light out of the scintillator to photomultipliers and achieves good longitudinal segmentation.

At larger pseudorapidities, closer to the beam pipe where higher radiation resistance is needed, liquid-argon technology is chosen for all calorimetry, for its intrinsic radiation tolerance. The hadronic endcap calorimeter is a copper/liquid-argon detector with parallel-plate

geometry, and the forward calorimeter is a dense liquid-argon calorimeter with rod-shaped electrodes in a tungsten matrix.

The approximately 200,000 signals from all of the liquid-argon calorimeters leave the cryostats through cold-to-warm feedthroughs located between the barrel and the extended barrel tile calorimeters, and at the back of each endcap. The barrel and extended barrel-tile calorimeters both support the liquid-argon cryostats and act as the flux return for the solenoid.

The CMS calorimeter system contrasts with that of ATLAS, because of its compactness[3,4]. In CMS, the solenoid is positioned outside the calorimeter, reducing the material in front of it, but also limiting the thickness of the calorimeter itself, and in particular the number of interaction lengths available to absorb hadronic showers.

The electromagnetic calorimeter, with coverage in pseudorapidity up to $|\eta| < 3.0$, comprises around 80,000 crystals of lead tungstate (Fig. 3b). These crystals have a high density and short radiation length, which make for a compact and high-resolution calorimeter. The main challenges are related to ageing/radiation effects and to temperature control (to the level of a tenth of a degree) to make full use the excellent intrinsic resolution of the system. The scintillation light is detected by silicon avalanche photodiodes in the barrel region and by vacuum phototriodes in the endcap region. A 'pre-shower' system, of silicon strip and lead layers, is installed in front of the endcaps to aid rejection of neutral pion signatures.

Surrounding the electromagnetic calorimeter is a brass/scintillator sampling hadron calorimeter, with coverage up to $|\eta| < 3.0$. Brass has a relatively short interaction length, is easy to machine and is non-magnetic. The scintillation light is converted by wavelength-shifting fibres embedded in the scintillator tiles and channelled to novel photodetectors known as hybrid photodiodes, which can operate in high axial magnetic fields.

Even with such compact electromagnetic and hadronic calorimeters in the barrel region, the total interaction length is limited to 7.2 λ (where λ is the interaction length or mean free path of a particle in the material) at $\eta = 0$ inside the solenoid coil. For this reason, a 'tail catcher' has been added around the coil to complement the hadronic calorimetry and to provide better protection against the escape (or 'punch-through') of hadronic energy into the muon system beyond.

The CMS forward calorimeter, constructed from steel and quartz fibres, is situated 11 m from the interaction point, thereby minimizing the amount of radiation and charge density in the detector during operation. The Čerenkov light emitted in the quartz fibres is detected by photomultipliers. The forward calorimeters ensure full geometric coverage up to $|\eta| = 5.0$, for the measurement of the transverse energy in the event and forward jet measurements.

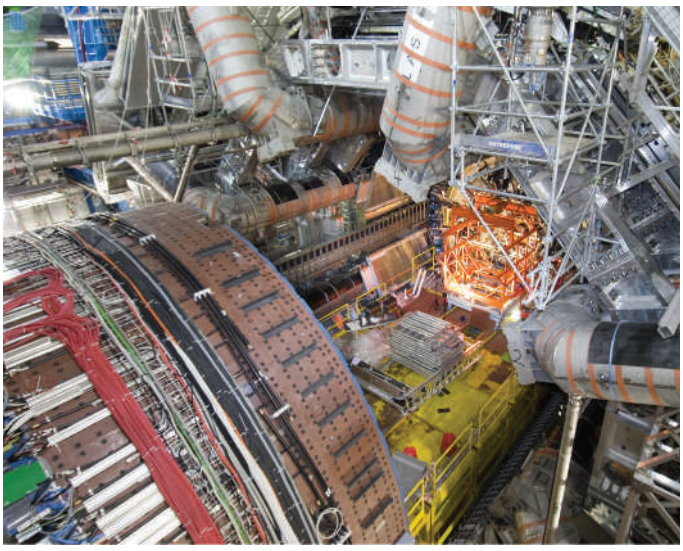


Figure 4 | Final integration. The components of the ATLAS detector are installed in the experiment's underground cavern. Here, part of the inner detector has just been moved inside the barrel calorimeter and toroid systems, while the endcap calorimeters (in the foreground) are kept in an open position to allow access. Image reproduced with permission from CERN.

Muon systems

The outermost detector layers of ATLAS and CMS, and the farthermost layers of LHCb, are dedicated to the measurement of the directions and momenta of high-energy muons, which escape from the calorimeters. Muons form a robust, clean and unambiguous signature of much of the physics of interest at the LHC.

Both ATLAS and CMS have had their overall detector designs optimized and adapted to trigger on and reconstruct muons at the highest luminosities of the LHC. (LHCb will operate, deliberately, at lower luminosity.) Muons must be measured with high efficiency and momentum resolution at low energies (such as in *B*-physics studies), at intermediate energies (for example, in the search for a Higgs decay into four muons), and at very high energies (to identify multi-TeV resonances such as a *Z*′). Wide pseudorapidity coverage is also important, and the ability to trigger on muons with energies of 5–10 GeV is crucial for several of the key physics goals. Finally, good timing resolution and the ability to identify in which proton-bunch crossing the muons were produced are absolute requirements, putting important constraints on the technological solutions chosen for the muon systems.

The muon systems in all three experiments are large-area gas-based detectors (several thousand square metres of multilayer chambers each in ATLAS and CMS). The chambers are divided into two sets, one intended for precise measurements of muon tracks and the other dedicated to triggering on muons. The sheer size of the systems means that there are significant technical challenges related to the stability and alignment of the chambers and to the careful mapping of the detectors' magnetic fields over large volumes. The radiation levels for the muon chambers are much less severe than for the inner detectors or calorimeters, but there are still concerns about ageing of the systems and also the neutron radiation environment of the experimental halls in which the detectors sit. The designs of the beam pipe and the shielding elements in the forward direction have been carefully optimized to reduce the neutron-induced background rates in the muon chambers.

Although the muon-chamber technologies chosen for ATLAS and CMS have many similarities, the magnet configuration in the two experiments is quite different. The ATLAS air-core toroid system, with a long barrel and two inserted endcap magnets, generates a large-volume magnetic field with strong bending power within a light and open structure. Multiple-scattering effects are thereby minimized, and excellent muon momentum resolution is achieved with three stations of high-precision muon-tracking chambers, covering up to $|\eta| = 2.7$. Over most of the range in pseudorapidity, the measurement of track coordinates (in the principal bending direction of the magnetic field) is performed by monitored drift tubes[1,2]. This technology provides robust and reliable operation, thanks to the mechanical isolation of each sense wire from its neighbours in the gas-filled drift volumes of the individual tubes. At large pseudorapidities and close to the interaction point, where the rate and background conditions are more difficult, cathode-strip chambers with higher granularity strip readout are used. The muon trigger system, with a fast time response and covering $|\eta| < 2.4$, comprises resistive-plate chambers in the barrel and thin gap chambers in the endcap regions. As well as triggering, these chambers provide a measurement of a second track coordinate orthogonal to the one measured by the high-precision chambers. In addition to the muon-chamber measurements, the inner detector measurements in the central solenoid of ATLAS contribute to the combined muon momentum resolution of the experiment.

The superconducting solenoid inside CMS is, at 13 m long with a 5.9-m inner diameter, the largest of its kind[3,4]. To achieve good momentum resolution without making overly stringent demands on muon-chamber resolution and alignment, the solenoid will operate at a high magnetic field of 4 T. In CMS, centrally produced muons are measured three times: in the inner tracker, after the coil, and in the return flux, into which four muon 'stations' are integrated, achieving robustness

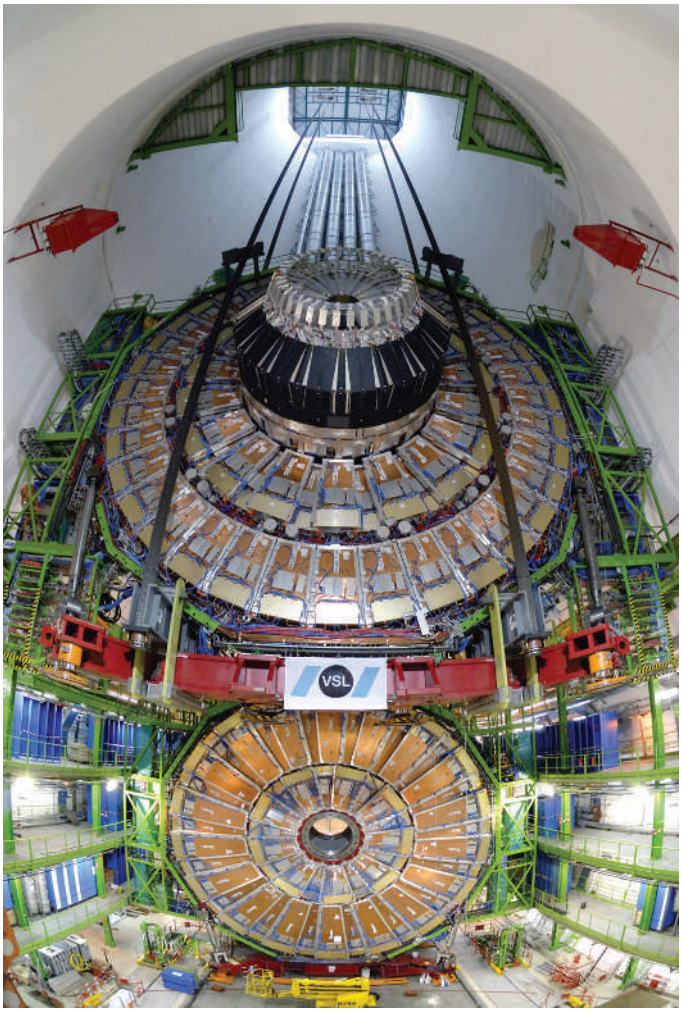


Figure 5 | Going underground. A large unit of the CMS detector, an endcap disc with muon chambers and part of the hadronic calorimeter, is lowered 100 m into its final position in the cavern. Image reproduced with permission from CERN.

and full geometric coverage. Three types of gaseous detector are used: drift tubes in the barrel region ($|\eta| < 1.2$), where the neutron-induced background is small, the muon rate is low and the residual magnetic field in the chambers is low; cathode-strip chambers in the two endcaps ($|\eta| < 2.4$), where the muon rate, the neutron-induced background and the magnetic field are all high; and resistive-plate chambers, in both the barrel and the endcap regions, to provide a fast response with good time resolution and to identify unambiguously the correct bunch crossing (albeit with coarser position resolution).

The intrinsic resolution of the high-precision chambers is in the range 60–150 μm (ref. 6), but the overall performance over the large areas involved (particularly at the highest momenta) depends on how well the muon chambers are aligned with respect to each other and with respect to the overall detector. The high accuracy of the ATLAS stand-alone muon measurement necessitates a precision of 30 μm on the alignment; in CMS, the different muon chambers need to be aligned with respect to each other and to the central tracking system to within 100–500 μm.

Both experiments have intricate hardware systems installed that are designed to measure the relative positions of chambers that contribute to the measurement of the same tracks, but also to monitor any displacements during the detector operation. For instance, in ATLAS, about 5,000 optical alignment sensors and 1,800 magnetic field sensors will track the movements of the chambers and will map and track the magnetic field to an accuracy of approximately 20 G (2 mT) throughout the detector volume. In the CMS, the solenoid magnetic field is more uniform but nevertheless the field is carefully monitored by around 80 sensors. Around 1,400 alignment sensors will provide independent monitoring of the tracking detector geometry with respect to an internal light-based reference system. The final alignment values will be obtained with the large statistics of muon tracks traversing the muon chambers.

The LHCb muon system consists of five stations. Multiwire proportional chambers are used throughout, except for the innermost region closest to the beamline of the first station. This station is placed in front of the calorimeters and represents a significant challenge in terms of material budget, space constraints, rate capability and radiation tolerance. The innermost region of this station, where the particle rates are highest, is equipped with triple-GEM (gas electron multiplier) detectors[13] with pad readout that are particularly suited for tracking in a high particle rate environment.

Triggering and readout

At design luminosity, the LHC will create 10^9 proton–proton events per second, but data storage and processing capabilities are such that data from only about 100–200 carefully selected events per second (each of these interesting events is accompanied by an average of 25 overlapping proton–proton events in the same bunch crossing) can be recorded offline for complete analysis. Hence, there is a need for a trigger system to select only the most important physics signatures and achieve a rejection factor of nearly 10^7.

The trigger systems for the LHC experiments have distinct levels. The first level, based on custom-built processors, uses a limited amount of the total detector information to make a decision in 2.5/3.2 μs (ATLAS/CMS) on whether to continue the processing of an event or not, reducing the data rate to around 100 kHz. Higher levels, using a network of several thousand commercial processors and fast switches and networks, access gradually more information and run algorithms that resemble offline data analysis to achieve the final reduction. The total amount of data recorded for each event will be roughly 1.5 megabytes, at a final rate of 150–200 Hz. This adds up to an annual data volume of the order of 10 petabytes for the LHC experiments.

The challenges to be faced in real-time data collection and reduction are many. The synchronization of the individual parts of the detector — and there are several thousand units to time in — must be accurate to better than a nanosecond, taking into account the flight times of particles to the individual sensor elements. At later stages, there is the second synchronization challenge of assembling all of the data for a particular bunch crossing from various parts of the detector into a complete event.

There is no chance of processing and selecting events within the 25 ns available between successive proton-bunch crossings. Furthermore, the sizes of the detectors and of the underground caverns in which they sit impose a minimum transit time between the detector electronics and trigger electronics. The first-level trigger calculations themselves need to be sufficiently sophisticated to identify clear physics signatures; the decision is based on the presence in the calorimeters or muon detectors of 'trigger primitive' objects, such as photons, electrons, muons and jets above pre-set transverse-energy or transverse-momentum thresholds. It also employs global sums of transverse energy and missing transverse energy. During the transit and processing time — less than 2.5/3.2 μs for ATLAS/CMS — the detector data must be time-stamped and held in buffers.

After an event is accepted by the first-level trigger system, the data from the pipelines are transferred from the detector electronics into readout buffers. The further processing involves signal processing, zero suppression and data compression while the events are examined by a farm of commodity processors consisting of several thousands of central processing units. The design and implementation of the processor farm, switching network, control software and trigger application software are major challenges. The event fragments must be directed from the readout buffers to a single processor and buffer node, using fast switches and networks, in order to perform more detailed calculations of the critical parameters of the event and to reduce the final rate further.

Even after such a large online reduction, huge amounts of data will be recorded. It was soon clear that the required level of computing resources

could be provided only by a significant number of computing centres working in unison with the CERN on-site computing facilities. Off-site facilities will be vital to the operation of the experiments to an unprecedented extent. Hence, over the past five years, GRID infrastructure for data processing and storage has been developed[14,15].

The GRID solution is geographically distributed and relies on three tiers[6]. The various tiers have clear responsibilities. The raw data output from the high-level trigger is processed and reconstructed in a Tier 0 computing facility at CERN, producing reconstructed data. Most detector and physics studies, with the exception of calibration and alignment procedures, will rely on this format. A copy of the raw data plus the reconstructed data is then sent to the Tier 1 centres around the world. These share the archiving of a second copy of the raw data, provide the reprocessing capacity and access to the various streams of reconstructed data (corresponding to the major trigger signatures) and allow data access and processing by the experiments' physics-analysis groups. The analysis data produced at Tier 0 and 1 are derived from the reconstructed data and are a reduced-event representation, intended to be sufficient for most physics analyses.

The Tier 2 facilities, each linked to a specific Tier 1 centre, are smaller but more numerous and are used for analysis, data-calibration activities and Monte Carlo simulations. Furthermore, the Tier 1 analysis data are copied to Tier 2 to improve access to them for physics analysis; by contrast, the Tier 1 centres provide safe storage of the large data sets produced at Tier 2 (for example, simulation data). A final level in the hierarchy is provided by individual group clusters and computers used for analysis.

This machinery for processing and analysis is being set up, and tests already indicate that the transfer speed between CERN and Tier 1 that is essential for initial running can be achieved, and that large-scale data production can be carried out in the GRID framework.

Ready to start

Installation of the LHC detectors in their underground caverns began in 2004. The components of ATLAS have been assembled and tested in the experiment's cavern, at 'Point 1' on the LHC ring (Fig. 4), and LHCb in the cavern at 'Point 8'. By contrast, the bulk of the CMS system was assembled and tested on the surface, before being lowered 100 metres into its cavern at 'Point 5' in 15 large lift operations (Fig. 5).

All of the LHC experiments have been operating their detector elements as much as possible on the surface and in 'test beams', and also, following installation, in the underground caverns. The flux of muons from cosmic rays provides a useful test of systems such as the calorimeters, inner detectors and muon systems to check alignment, calibration and the integration of data collection. In parallel, the GRID computing infrastructure and organisation are being planned, implemented and tested. For ATLAS, CMS and LHCb, major exercises of their data processing, software and computing infrastructure have been performed, and more are planned in the run-up to the introduction of beams into the LHC in 2008.

The first collisions at a centre-of-mass energy of 14 TeV are expected by mid-2008. As soon as a luminosity of 10^{32} $cm^{-2}s^{-1}$ is reached, within days the LHC can produce data sets — of W and Z bosons, t quarks, high-transverse-momentum jets, and even supersymmetric particles — that will surpass those of any previous or existing accelerator. At that point, the main physics goals of the LHC will be in full focus, and the aim will be to collect as much data in 2008 as possible.

The decade-long period of detector development and construction is coming to an end. Many of the fundamental challenges addressed by the experiment builders at LHC have been solved successfully, most notably in the development of fast, radiation-tolerant and sufficiently granular detector systems and electronics, integrated in turn in large systems that exceed any existing detector in specification (size, speed and channel count, in particular). Technology advances in computing, switches, networks and software have allowed the development of sophisticated trigger, data acquisition and GRID systems to handle the LHC data rates and volumes — it was far from obvious that this could be achieved when the building of the experiments was initially approved. The accelerator, detectors and off-line systems now need to be completed in their underground areas, commissioned fully and operated efficiently.

A further challenge overcome in the LHC project is the successful collaboration in each experimental team of as many as 2,000 scientists from all over the world, working together for a decade and using their resources and skills efficiently. Thanks to their efforts, and those of the teams building the LHC itself, a new era of research in experimental particle physics is finally within reach. The community can now look forward to the new challenges posed in interpreting the data from the LHC — challenges as great as those that have been faced in the building of the detectors. There is good reason to be optimistic, and the potential rewards, in terms of physics discoveries, make it well worth the effort. ■

1. ATLAS Collaboration. *ATLAS Detector and Physics Performance: Technical Design Report*. Report No. CERN/LHCC/99-15 (CERN, Geneva, 1999).
2. ATLAS Collaboration. *ATLAS: Technical Proposal for a General-Purpose p p Experiment at the Large Hadron Collider at CERN*. Report No. CERN/LHCC/94-43 (CERN, Geneva, 1994).
3. CMS Collaboration. *CMS Physics Technical Design Report* Vol. I. Report No. CERN/LHCC/2006-01 (CERN, Geneva, 2006).
4. CMS Collaboration. *CMS Technical Proposal*. Report No. CERN/LHCC/94-38 (CERN, Geneva, 1994).
5. Ellis, N. & Virdee, T. S. Experimental challenges in high luminosity collider physics. *Annu. Rev. Nucl. Part. Sci.* **44,** 609–653 (1994).
6. Froidevaux, D. & Sphicas, P. General purpose detectors for the Large Hadron Collider. *Annu. Rev. Nucl. Part. Sci.* **56,** 375–440 (2006).
7. Gianotti, F. *European School of High-Energy Physics, CERN Yellow Report*. Report No. CERN-2000-007 219–244 (CERN, Geneva, 1999).
8. LHCb Collaboration. *LHCb Technical Proposal*. Report No. CERN/LHCC/98-104 (CERN, Geneva, 1998).
9. Yao, W.-M. *et al*. Review of particle physics. *J. Phys. G* **33,** 1 (2006).
10. Doležal, Z. *et al*. The silicon microstrip sensors of the ATLAS semiconductor tracker. *Nucl. Instrum. Methods* **578,** 98–118 (2007).
11. CMS Collaboration. *Addendum to the CMS Tracker TDR*. Report No. CERN/LHCC 2000-016 (CERN, Geneva, 2000).
12. LHCb Collaboration. *LHCb TDR 5* Report No. CERN/LHCC 2001-011 (CERN, Geneva, 2001).
13. LHCb Collaboration. *Second Addendum to the Muon System Technical Design Report*. Report No. CERN/LHCC/2005-0012 (CERN, Geneva, 2005).
14. Foster, I. & Kesselman, C. *The Grid: Blueprint for a New Computing Infrastructure* (Morgan & Kaufmann, San Francisco, 1998).
15. LCG project. *LHC Computing Grid Technical Design Report*. Report No. CERN/LHCC/2005-024 (CERN, Geneva, 2005).

Acknowledgements This review article is largely based on the very complete documentation already existing for the ATLAS, CMS and LHCb experiments, and on refs 1–8 in particular. Furthermore, with around 5,000 scientists involved in the construction of the experiments described, I can only cover a small part of the challenges, excitement and difficulties involved in making them a reality. The best I can do is therefore to acknowledge all of the members of these collaborations, and in particular those who have helped me with corrections and comments, and apologize for all the dedicated sub-projects and work I had to leave out of this review.

Author information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (steinar.stapnes@cern.ch).

Beyond the standard model with the LHC

John Ellis[1]

Whether or not the Large Hadron Collider reveals the long-awaited Higgs particle, it is likely to lead to discoveries that add to, or challenge, the standard model of particle physics. Data produced will be pored over for any evidence of supersymmetric partners for the existing denizens of the particle 'zoo' and for the curled-up extra dimensions demanded by string theory. There might also be clues as to why matter dominates over antimatter in the Universe, and as to the nature of the Universe's dark matter.

The unparalleled high energy of the Large Hadron Collider (LHC), with its 7 TeV per beam and its enormously high collision rate that should reach a billion collisions per second, makes it a microscope able to explore the inner structure of matter on a scale that is an order of magnitude smaller than previously achieved. Results at the energies and distances explored so far led physicists to successfully describe matter using the standard model of particle physics[1–3]. But this description is incomplete, and the standard model raises, but leaves unanswered, many fundamental questions. Explanations are needed for the origin of particle masses and the small differences seen in the properties of matter and antimatter, as well as to establish whether fundamental interactions can be unified. Moreover, the standard model has no explanation for some of the basic puzzles of cosmology, such as the origin of matter and the nature of the Universe's dark matter and dark energy. There are high hopes that the LHC will help resolve at least some of these basic issues in cosmology and in physics beyond the standard model[4].

Theoretical calculations made using the standard model agree well with data collected at lower-energy accelerators, such as at CERN's Large Electron–Positron (LEP) accelerator in the 1990s and, more recently, at the Tevatron proton–antiproton collider at Fermilab (Batavia, Illinois)[5]. Data collected at LEP agreed with the standard model at the per-mille level, and recent measurements of the masses of the intermediate vector boson W (ref. 6) and the top quark[7] agree well with standard-model predictions. But the theoretical calculations are valid only with an ingredient that has not yet been observed — the notorious Higgs boson. Without this missing ingredient, the calculations yield incomprehensible, infinite results[8,9]. The agreement of the data with the calculations implies not only that the Higgs boson (or something equivalent) must exist, but also suggests that its mass should be well within the reach of the LHC[5].

In this review, I discuss the likelihood of finding the Higgs boson and what other physics beyond the standard model the accelerator might reveal.

Searching for symmetry breaking

Why should the Higgs boson exist, and are there any alternatives? In the underlying equations of the standard model, none of the elementary particles seems to have mass. In the real world, however, only the photon and gluon, the carriers of the electromagnetic and strong nuclear interactions, are massless. All the other elementary particles are massive, with the W and Z bosons, intermediaries of the weak nuclear interaction, and the top quark weighing as much as decent-sized nuclei. The underlying symmetry between the different particles of the standard model must be broken so that some may acquire masses.

There are two ways to break the symmetry of the standard model. The preferred way is to respect the symmetry of the underlying equations, in which the massless photon and the massive W and Z bosons appear in the same way, but look for an asymmetric solution, much as the reader and writer are lopsided solutions of the symmetric equations of electromagnetism. According to this approach to the standard model, symmetry is thought to be already broken in the lowest-energy state, the so-called vacuum. This 'spontaneous' symmetry breaking is ascribed to a field that permeates all space, taking a specific value that can be calculated from the underlying equations, but with a random orientation in the internal 'space' of particles that breaks the underlying symmetry. This mechanism, which was suggested by Peter Higgs[10] and independently by Robert Brout and François Englert[11], forces some particles, such as the photon, to remain massless, but gives masses to others in proportion to their coupling to this vacuum field (Fig. 1).

In the same way that the electromagnetic field has a quantum particle associated with it, the photon, this vacuum field would also have an associated quantum particle, the Higgs boson. Experiments at LEP seemed at one time to have found a hint of its existence[12]. In the end, however, these searches were unsuccessful and told us only that any Higgs boson must weigh at least 114 GeV (ref. 13). If its mass is less than about 200 GeV, researchers using the Tevatron may find some evidence for it before the LHC comes into operation[14].

The large experiments, ATLAS[15] and CMS[16], at the LHC will be looking for the Higgs boson in several ways (Fig. 2). The Higgs boson is predicted to be unstable and decay into other particles, such as photons, bottom quarks, tau leptons, W or Z bosons. It may well be necessary to combine several different decay modes to uncover a convincing signal. The LHC experiments should be able to find the Higgs boson even if it weighs as much as 1 TeV, and there are high expectations that it could be found during the first couple of years of LHC operation. Its discovery would set the seal on the success of the standard model.

Higgs or bust?

With the impending confirmation or refutation of the Higgs hypothesis, many theorists are getting cold feet. Some are beginning to support alternative scenarios that go beyond the standard model[17]. One popular suggestion is that the Higgs boson might not be an 'elementary' particle in the same sense as the quarks, leptons and the photon, but instead might be composed of simpler constituents[18]. This model would be analogous to the Bardeen–Cooper–Schrieffer (BCS) theory of superconductivity, in which a photon acquires an effective mass by interacting with 'Cooper pairs' of electrons. In this analogy, the W and Z bosons would 'eat' tightly bound pairs of novel strongly interacting fermions

[1]Theory Division, Physics Department, CERN, CH-1211 Geneva 23, Switzerland.

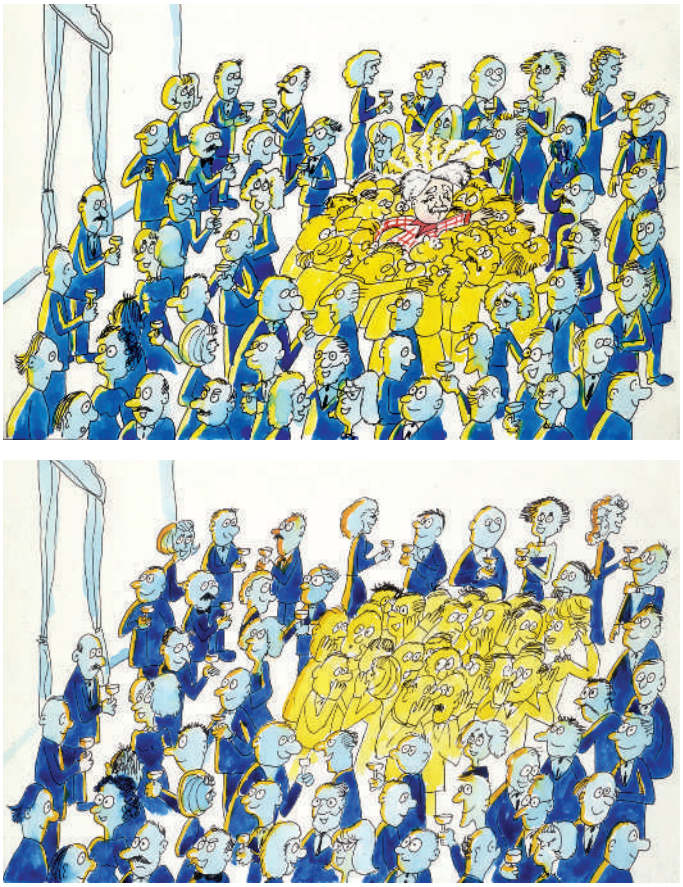


Figure 1 | Picturing the Higgs field. The behaviour of physicists in a crowded social event at a conference is an analogy for the Higgs mechanism, as proposed by David Miller (University College London). The physicists represent a non-trivial medium permeating space. In the upper panel, the physicists cluster around a famous scientist who enters the room, slowing the scientist's progress. In much the same way, a particle passing through the Higgs–Brout–Englert field slows down and acquires a mass. In the lower panel, a rumour propagates. This is an excitation of the medium — the group of physicists — itself, forming a body with a large mass; this is analogous to the formation of a Higgs boson. Figure reproduced with permission from CERN.

rather than an elementary Higgs field. It seems rather difficult to reconcile this composite alternative with the accurate low-energy data from LEP[5], but some enthusiasts are still pursuing this possibility. Alternatively, it has been suggested that the Higgs boson is indeed elementary, but is supplemented by some additional physics — for example, being supersymmetric (discussed later).

The most radical alternative to the Higgs hypothesis exploits the second way of breaking the standard model's symmetry. It postulates that, although the underlying equations are symmetric, their solution is subject to boundary conditions that break that symmetry. What boundary would that be, given that space is apparently infinite (or at least very large compared to the scale of particle physics)? The answer is that there might be additional, very small dimensions of space with edges where the symmetry may be broken[19]. Such models would have no Higgs boson, and are difficult to reconcile with the data already acquired that seem to require a relatively light Higgs boson.

Theorists are amusing themselves discussing which would be worse: to discover a Higgs boson with exactly the properties predicted in the standard model or to discover that there is no Higgs boson. The former would be a vindication of theory, but would teach us little new. The latter would upset the entire basis of the standard model. The absence of a Higgs boson would be exciting for particle physicists, but it might not be so funny to explain to the politicians who have funded the LHC mainly to discover this particle. Whichever option nature chooses, the good news is that the LHC will provide us with a clear-cut experimental answer and end the speculation.

The hierarchy problem

Resolving the Higgs question will set the seal on the standard model, but, as I mentioned at the beginning, there are plenty of reasons to expect other physics beyond the standard model to be discovered (Fig. 3). Specifically, there are good reasons to expect other discoveries at the TeV energy scale, within reach of experiments at the LHC. Many would consider this to be the primary motivation for the leap into the unknown that the LHC represents.

For example, it is generally thought that the elementary Higgs boson of the standard model cannot exist in isolation. Specifically, difficulties arise when one calculates quantum corrections to the mass of the Higgs boson owing to the exchanges of virtual particles (see, for example, ref. 20). Not only are these corrections infinite in the standard model, but, if the usual procedure of controlling them by cutting the theory off at some high energy or short distance is adopted, the net result depends on the square of the cut-off scale. This implies that, if the standard model were embedded in some more complete theory that kicks in at high energy — such as a grand unified theory of the particle interactions or a quantum theory of gravity — the mass of the Higgs boson would be sensitive to the details of this high-energy theory. This would make it difficult to understand why the Higgs boson has a (relatively) low mass. It would also, by extension, make it difficult to explain why the energy scale of the weak interactions — as reflected in the masses of the W and Z bosons — is so much smaller than that of unification or quantum gravity.

One might be tempted simply to wish away this 'hierarchy problem' by postulating that the underlying parameters of the theory are tuned finely, so that the net value of the Higgs boson mass obtained after adding in the quantum corrections is unnaturally small as the result of some sneaky cancellation. But it would surely be more satisfactory either to abolish the extreme sensitivity to the quantum corrections or to cancel them in a systematic manner. Indeed, this has been one of the reasons for believing that the Higgs boson is composite. If it is, the Higgs boson would have a finite size, which would cut the pesky quantum corrections off at some relatively low scale. In this case, the LHC might uncover a cornucopia of new particles with masses around this cut-off scale, which should be near 1 TeV. At the very least, the interactions of the W and Z vector bosons would be modified in an observable way.

The supersymmetric solution

An alternative way to get rid of these quantum corrections is provided by supersymmetry[21]. This is an elegant theory that would pair up fermions, such as the quarks and leptons that make up ordinary matter, with bosons, such as the photon, gluons, W and Z that carry forces between the matter particles or even the Higgs itself (Fig. 4). Supersymmetry also seems to be essential for making a consistent quantum theory of gravity based on string theory (of which more later). However, these elegant arguments give no clue as to what energies would be required to observe supersymmetry in nature.

The first argument that supersymmetry might appear near the TeV scale was provided by the hierarchy problem: in a supersymmetric theory, the quantum corrections owing to the pairs of virtual fermions and bosons cancel each other systematically[22], and a low-mass Higgs boson no longer seems unnatural[23]. The residual quantum corrections to the mass of the Higgs boson would be small if differences in mass between supersymmetric partner particles were less than about 1 TeV. Because the fermions and bosons of the standard model do not pair up with each other in a neat supersymmetric manner, this theory would require each of the standard-model particles to be accompanied by an as-yet unseen supersymmetric partner. It might seem profligate for there to be all these partners, but at least the hypothesis predicts a 'scornucopia' of supersymmetric particles that should weigh less than about 1 TeV and hence could be produced by the LHC[15,16].

In the wake of this hierarchy argument, at least three other reasons have surfaced for thinking that supersymmetric particles weigh about 1 TeV.

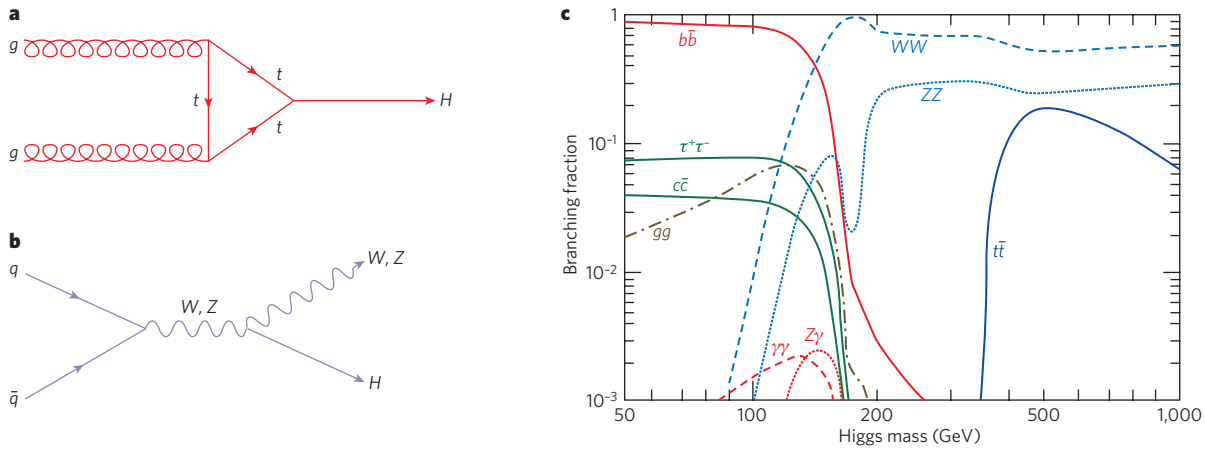


Figure 2 | The Higgs boson at the LHC. A Higgs (H) boson may be produced by a range of interactions, two examples of which are shown here. The first, **a**, is through fusion of gluons (g) from the protons in the LHC beams, through a top (t) quark loop; and the second, **b**, is through a *bremsstrahlung* process, in which a quark (q) and antiquark ($\bar{q}$) annihilate to create a W or Z boson, which may then radiate a Higgs. **c**, The Higgs itself then decays, and it is these decay products that will be caught in a detector. The 'branching fraction' or probability of decay to certain products depends on the (as-yet unknown) mass of the Higgs particle, which is dominated by decay to a bottom–antibottom quark pair at low mass, but by decay to pairs of W bosons at high mass.

The first is that these particles would facilitate the unification of the strong, weak and electromagnetic forces into a simple grand unified theory[24]. Another argument is that a theory with low-energy supersymmetry would predict that the Higgs boson weighs less than about 150 GeV (ref. 25), which is precisely the range favoured indirectly by the present data. The final one is that, in many models, the lightest supersymmetric particle (LSP) is an ideal candidate for the dark matter advocated by astrophysicists and cosmologists.

The LSP is ideal because it is stable when a suitable combination of baryon and lepton numbers is conserved[26], as happens in the minimal supersymmetric extension of the standard model, as well as in simple models of grand unification and neutrino masses. In this case, LSPs would be left over as relics from early in the Big Bang, and calculations of their abundance yield a density of dark matter in the range favoured by astrophysics and cosmology if the LSP weighs at most a few hundred GeV, probably putting it within reach of the LHC[27].

Supersymmetry could be a bonanza for the LHC, with many types of supersymmetric particle being discovered. In many models, the LHC would produce pairs of gluinos (the supersymmetric partners of the gluons) or squarks (the supersymmetric partners of the quarks) that would subsequently decay through various intermediate supersymmetric particles. Finally, each of these pairs of particles would yield a pair of LSPs that interact only weakly and hence carry energy away invisibly. In favourable cases, the masses of several intermediate particles could be reconstructed this way. It might even be possible to use these measurements to calculate what the supersymmetric dark-matter density should be, so as to compare the result with the astrophysical estimates[28].

Into extra dimensions?

Postulating a composite Higgs boson or supersymmetry are not the only strategies that have been proposed for dealing with the hierarchy problem. Another suggestion is that there are additional dimensions of space[29]. Clearly, space is three-dimensional on the scales that we know so far, but the idea that there are additional dimensions curled up so small that they are invisible has been in the air since it was first proposed by Kaluza and Klein over 80 years ago. This idea has gained ground in recent years with the realization that string theory predicts the existence of extra dimensions of space[30].

According to string theory, elementary particles are not idealized points of euclidean geometry, but are objects extended along one dimension (a string) or are membranes with more dimensions[31]. For the quantum theory of strings to be consistent, particles have to move in a space with more than the usual three dimensions. Initially, it was thought that these extra dimensions would be curled up on scales that might be as small as the Planck length of around 10^{-33} cm. But more recently, it was realized that at least some of these new dimensions might be much larger and possibly have consequences observable at the LHC.

One of the possibilities offered by these speculations is that gravity is strong when these extra dimensions appear, possibly at energies close to 1 TeV. Under this condition, according to some variants of string theory, microscopic black holes might be produced by the LHC[32]. These would be short-lived, decaying rapidly through thermal (Hawking) radiation. Measurements of this radiation would offer a unique laboratory window on the mysteries of quantum gravity. The microscopic black holes would emit energetic photons, leptons, quarks and neutrinos, providing distinctive experimental signatures. In particular, the neutrinos they emit would carry away more invisible energy than LSPs would in the supersymmetric models discussed previously[33].

Although microscopic black holes would be the most dramatic sign of large extra dimensions, they are not the only sign of such theories that might be visible at the LHC. If the extra dimensions are curled up on a sufficiently large scale, the ATLAS and CMS projects might be able to see Kaluza–Klein excitations of standard-model particles, or even of the graviton, the mediator particle of gravity. Indeed, the spectroscopy of some extra-dimensional theories might be as rich as that of supersymmetry[34]. If so, how do we tell which cornucopia the LHC is uncovering? There are significant differences in the relationship between, for example, the masses of the partners of quarks and leptons in supersymmetric theories and in theories with large extra dimensions. Moreover, the spins of the Kaluza–Klein excitations would be the same as those of their standard-model progenitors, whereas the spins of the supersymmetric partners

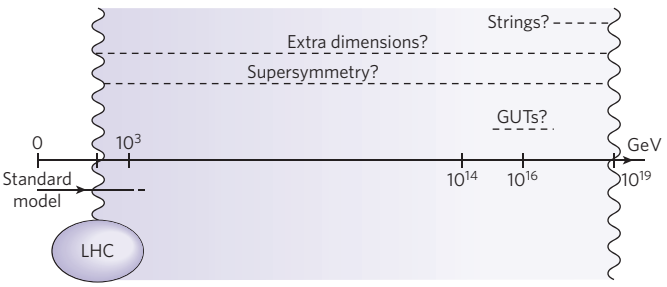


Figure 3 | Physics beyond the TeV scale. The standard model has been well tested up to around the 100-GeV mass scale. The LHC will test beyond this, to the crucial 1,000-GeV level, the TeV scale, at which hints of new physics, such as supersymmetry and extra dimensions, may emerge. String theory or grand unified theories (GUTs) inhabit much higher energy scales, approaching 10^{19} GeV, which is called the Planck scale.

	Known particles of the standard model	Postulated supersymmetric partners or 'sparticles'	
Half-integer spin	Electron	Selectron	Integer spin
	Neutrino	Sneutrino	
	Top quark	Stop	
Integer spin	Gluon	Gluino	Half-integer spin
	Photon	Photino	

Figure 4 | Examples of supersymmetric partners. Supersymmetry is a symmetry drawn between fermions (with half-integer spin) and bosons (with integer spin). It postulates that, for each fermion, there exists a bosonic partner — such as the supersymmetric electron, or 'selectron', which partners the electron. Similarly, each boson is thought to have a fermionic superpartner, which for the gluon is the 'gluino'.

would be different. These underlying differences translate into characteristic differences in the spectra of decay products in the two classes of model and into distinctive correlations between them[35].

It is amusing that, in some theories with extra dimensions, the lightest Kaluza–Klein particle (LKP) might be stable[36], rather like the LSP in supersymmetric models. In this case, the LKP would be another candidate for astrophysical dark matter. Thus, there is more than one way in which LHC physics beyond the standard model might explain the origin of dark matter: fortunately, the tools seem to be available for distinguishing between them.

The matter–antimatter conundrum

Will the LHC explain the origin of conventional matter? As was first pointed out by the Russian physicist Andrei Sakharov[37], particle physics can explain the origin of matter in the Universe in terms of small differences in the properties of matter and antimatter, such as those discovered in the decays of *K* and *B* mesons. Present experimental data accord well with the matter–antimatter differences allowed by the standard model. However, by themselves, these differences in the properties of matter and antimatter would be insufficient to generate the matter seen in the Universe. It is possible that the deficit will be explained by new physics at the TeV scale revealed by the LHC. For example, supersymmetry allows many more possibilities for differences between the properties of matter and antimatter than are possible in the standard model[38]; some of these differences might explain the amount of matter in the Universe.

This provides one of the motivations for the LHCb experiment[39], which is dedicated to probing the differences between matter and antimatter, notably looking for discrepancies with the standard model (Box 1). In particular, LHCb has unique capabilities for probing the decays of mesons containing both bottom and strange quarks, the constituents of the *B* and *K* mesons probed in other experiments investigating matter–antimatter differences. There are many other ways to explore the physics of matter and antimatter, and the ATLAS and CMS experiments will also contribute to them, in particular by searching for rare decays of mesons containing bottom quarks.

If these experiments detect any new particles beyond the standard model at the TeV scale, questions will immediately arise as to whether this new physics distinguishes between matter and antimatter, and whether or not this new physics explains the origin of matter in the Universe. For example, if the Higgs boson is discovered at the LHC, are its couplings to matter and antimatter the same? If supersymmetry is discovered at the LHC, do supersymmetric 'sparticles' and 'antisparticles' behave in the same way? There are many models in which matter–antimatter differences in the Higgs or sparticle sector are responsible for the origin of the matter in the Universe.

Into the future

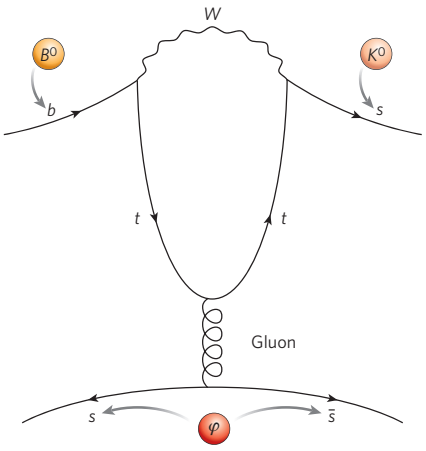
According to present plans, the first full-energy collisions of the LHC will take place in 2008, although it will take some time for the accelerator to build up to its designed nominal collision rate. There are hopes, however, that in its first couple of years of operation, it will already start to provide crucial information on physics beyond the standard model, for example by discovering the Higgs boson — or other new particles such as those

Box 1 | Penguin hunting at LHCb

Why does matter dominate over antimatter in the Universe, considering that both were thought to be created in equal quantities in the Big Bang? Part of the explanation is that some interactions between particles take place at different rates when two fundamental symmetries of the quantum field theory that underlie them are simultaneously reversed. There are two symmetries involved: charge conjugation, C; and parity symmetry, P. Charge conjugation turns particles into their antiparticles by reversing internal properties such as electric charge. By contrast, parity symmetry flips external particle properties such as spin, similar to looking at an interaction in a mirror.

CP violation was first discovered experimentally[45] in decays of *K* mesons, which contain a strange quark in addition to an up or down quark. Later theoretical work showed[46] that CP violation would occur naturally in interactions mediated by the weak nuclear force in the standard model with three quark generations. (At the time, particles from only two were known.) The degree of violation is, however, insufficient to explain the Universe's matter–antimatter imbalance.

The subsequent discovery of the third quark generation, formed of the bottom (*b*) and top (*t*) quarks, vindicated the model. A plethora of



experiments has since confirmed CP violation, indirectly and directly, in decay channels of *B* mesons (those containing bottom quarks), where the effect is expected to be particularly large. These experiments notably include two specially constructed '*B* factories', the Belle detector at KEK in Japan, and BaBar at the Stanford Linear Accelerator Center (SLAC), in California, which have delivered a series of more precise values for the parameters of CP violation since 2001.

The LHCb experiment is LHC's dedicated CP-violation detector. It is a 20-m-long spectrometer with a conical detection volume expanding in radius along the beam axis. It is attuned to detecting the distinctive signature of *B* decays — charged particles with high transverse momenta originating from a vertex significantly displaced from the interaction point of the proton beams.

To maximize the probability of only a single *B* interaction per beam crossing, the LHC beams are defocused slightly to a luminosity of around $2.5 \times 10^{32}\,cm^{-2}s^{-1}$, below the LHC's nominal value of $10^{34}\,cm^{-2}s^{-1}$. The implied collision rates and the high energy of the LHC beams should allow CP-violation parameters to be more tightly constrained and perhaps also provide a glimpse of physics beyond the standard model.

Such physics could manifest itself, in particular, in 'penguin' processes such as $B^0 \rightarrow K^0 \varphi$ (see page 270) in which the decay of a highly energetic *B* meson takes place, legitimately according to the rules of the weak interaction, through an intermediate loop of massive particles such as a top quark and a *W* boson (see figure). Does the degree of CP violation in such a process differ significantly from that found, for example, in the decay $B^0 \rightarrow K^0 J/\Psi$, which does not include a penguin loop? If so, that could be an indication of new physics participating in the penguin loop — such as the involvement of supersymmetric particles.

predicted by supersymmetry, if they are not too heavy[40]. Continued running of the LHC at its nominal luminosity would enable many properties of the Higgs boson to be verified, for example by providing measurements of its couplings to some other particles and checking whether these are proportional to the particles' masses. This period should also enable the properties of any other newly discovered particles to be checked, such as establishing whether their spins are the same as those of their standard-model counterparts or are different.

What might be possible using the LHC after these planned phases of exploitation? One possibility is to add new components to the existing ATLAS and CMS detectors that would provide new ways to study the Higgs boson. For example, new components close to the beams several hundred metres from the interaction points might be able to detect rare proton–proton collisions that produce nothing except a single isolated Higgs boson[41]. Another possibility is that supersymmetric or other new particles might show up in unexpected ways. For example, in some supersymmetric scenarios there would be a metastable charged particle that would have quite distinctive experimental signatures[42], and it might be interesting to devise new detectors to explore this possibility.

It might also be possible to increase the LHC collision rate significantly beyond the nominal value. This possibility would be particularly interesting if, for example, the initial runs of the LHC discover new physics with a very low production rate, perhaps because it has a high energy threshold. Increasing the LHC collision rate might be possible by redesigning the collision points using new magnet technologies; it would also require replacing at least some of CERN's lower-energy accelerators, such as the low-energy linear proton accelerator and the Proton Synchrotron, so as to feed more intense beams into the LHC[43]. Technical options for increasing the LHC collision rate are now being evaluated, so that they can be considered when the first experimental results from the initial LHC runs become available, some time around 2010.

Exploitation of the LHC and the study of possible upgrade options are among the highest priorities for European particle physics and were decided upon at a special meeting of the CERN Council in Lisbon in July 2006 (ref. 44). Possible future accelerators were also considered, such as a linear electron–positron collider or a neutrino factory. The priorities for these options will surely depend on the nature and energy scale of whatever new physics beyond the standard model the LHC reveals, as well as on developments in other areas such as neutrino physics. A central element in the European strategy for particle physics is the need to review advances in particle physics in the coming years, and in particular to review the implications of any LHC discoveries at the end of this decade.

Particle physics stands on the brink of a new era. Research using the LHC will make the first exploration of physics in the TeV energy range. There are good reasons to hope that the LHC will find new physics beyond the standard model, but no guarantees. The most one can say for now is that the LHC has the potential to revolutionize particle physics, and that in a few years' time we should know what course this revolution will take. Will there be a Higgs boson, or not? Will space reveal new properties at small distances, such as extra dimensions or supersymmetry? Will experiments at the LHC cast light on some fundamental cosmological questions, such as the origin of matter or the nature of dark matter? Whatever the answers to these questions might be or whatever surprises the LHC might spring, it will surely set the agenda for the next steps in particle physics.■

1. Glashow, S. Partial symmetries of weak interactions. *Nucl. Phys.* **22,** 579–588 (1961).
2. Weinberg, S. A model of leptons. *Phys. Rev. Lett.* **19,** 1264–1266 (1967).
3. Salam, A. in *Elementary Particle Physics: Relativistic Groups and Analyticity* (Nobel Symposium No. 8) (ed. Svartholm, N.) 367 (Almqvist and Wiksills, Stockholm, 1968).
4. Ellis, J. Physics at LHC. *Acta Phys. Polon.* **B38,** 1071–1092; preprint at <http://arxiv.org/abs/hep-ph/0611237> (2007).
5. LEP Electroweak Working Group. *LEP Electroweak Working Group* <http://lepewwg.web.cern.ch/LEPEWWG/> (2007).
6. Tevatron Electroweak Working Group. *Tevatron Electroweak Working Group W/Z Subgroup* <http://tevewwg.fnal.gov/wz/> (2007).
7. Tevatron Electroweak Working Group. A combination of CDF and D0 results on the mass of the top quark. Preprint at <http://arxiv.org/abs/hep-ex/0703034> (2007).
8. 't Hooft, G. Renormalizable Lagrangians for massive Yang–Mills fields. *Nucl. Phys. B* **35,** 167–188 (1971).
9. 't Hooft, G. & Veltman, M. Regularization and renormalization of gauge fields. *Nucl. Phys. B* **44,** 189–213 (1972).
10. Higgs, P. W. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.* **13,** 508–509 (1964).
11. Englert, F. & Brout, R. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.* **13,** 321–322 (1964).
12. Barate, R. *et al.* (ALEPH Collaboration) Observation of an excess in the search for the standard model Higgs boson at ALEPH. *Phys. Lett. B* **495,** 1–17 (2000).
13. LEP Higgs Working Group. *LEP Higgs Working Group* <http://lephiggs.web.cern.ch/LEPHIGGS/www/Welcome.html> (2007).
14. Cavalli, D. *et al.* The Higgs working group: summary report. *Proc. Workshop on Physics at TeV Colliders* (Les Houches, 2001).
15. ATLAS Collaboration. *Detector and Physics Performance Technical Design Report.* <http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/TDR/access.html> (2007).
16. CMS Collaboration. *CMS Physics: Technical Design Report* (ed. De Roeck, A.) CERN-LHCC-2006-021 (2006).
17. Ellis, J. Summary of the International Conference on High-Energy Physics, Beijing, China, August 2004. *Int. J. Mod. Phys.* **A20**, 5297 (2005).
18. Farhi, E. & Susskind, L. Technicolor. *Phys. Reports* **74,** 277–321 (1981).
19. Csaki, C., Grojean, C. Pilo, L. & Terning, J. Towards a realistic model of Higgsless electroweak symmetry breaking. *Phys. Rev. Lett.* **92,** 101802 (2004).
20. 't Hooft, G. Naturalness, chiral symmetry, and spontaneous chiral symmetry breaking. Under the spell of the gauge principle. (eds 't Hooft, G. *et al.*) 352–374 (World Scientific, Singapore, 1994).
21. Wess, J. & Zumino, B. A Lagrangian model invariant under supergauge transformations. *Phys. Lett. B* **49,** 52–54 (1974).
22. Ferrara, S., Iliopoulos, J. & Zumino, B. Supergauge invariance and the Gell-Mann–Low eigenvalue. *Nucl. Phys. B* **77,** 413–419 (1974).
23. Witten, E. Mass hierarchies in supersymmetric theories. *Phys. Lett. B* **105,** 267–271 (1981).
24. Ellis, J., Kelley, S. & Nanopoulos, D. *Phys. Lett. B* **249,** 441–448 (1990).
25. Ellis, J., Ridolfi, G. & Zwirner, F. Higgs boson properties in the standard model and its supersymmetric extensions. Preprint at <http://arxiv.org/pdf/hep-ph/0702114> (2007).
26. Fayet, P. in *Unification of the Fundamental Particle Interactions* (eds Ferrara, S., Ellis, J.& van Nieuwenhuizen, P.) 587 (Plenum, New York, 1980).
27. Ellis,J., Hagelin, J., Nanopoulos, D., Olive, K. & Srednicki, M. Supersymmetric relics from the Big Bang. *Nucl. Phys. B* **238,** 453–476 (1984).
28. Battaglia, M. *et al.* Updated post-WMAP benchmarks for supersymmetry. *Eur. Phys. J.* **C33,** 273–296 (2004).
29. Antoniadis, I. A possible new dimension at a few TeV. *Phys. Lett. B* **246,** 377–384 (1990).
30. Green, M., Schwarz, J. & Witten, E. *Superstring Theory* (Cambridge Univ. Press, Cambridge, 1987).
31. Randall, S. & Sundrum, R. An alternative to compactification. *Phys. Rev. Lett.* **83,** 4690–4693 (1999).
32. Arkani-Hamed, N., Dimopoulos, S. & Dvali, G. The hierarchy problem and new dimensions at a millimeter. *Phys. Lett. B* **429,** 263–272 (1998).
33. Harris, C. M. *et al.* Exploring higher dimensional black holes at the Large Hadron Collider. *JHEP* **0505,** 053 (2005).
34. Cembranos, J., Feng, J. Rajaraman, A. & Takayama, F. Exotic collider signals from the complete phase diagram of minimal universal extra dimensions. *Phys. Rev. D* **75,** 036004 (2007).
35. Athanasiou, C., Lester, C. G., Smillie, J. M. & Webber, B. R. Distinguishing spins in decay chains at the Large Hadron Collider. *JHEP* **0608,** 055 (2006).
36. Servant, G. & Tait,T. M. P. Is the lightest Kaluza–Klein particle a viable dark matter candidate? *Nucl. Phys. B* **650,** 391–419 (2003).
37. Sakharov, A. D. Violation of CP invariance, C asymmetry, and baryon asymmetry of the Universe. *Pisma Zh. Eksp. Teor. Fiz.* **5,** 32–35 (1967).
38. Cline, J., Joyce, M. & Kainulainen, K. Supersymmetric electroweak baryogenesis. *JHEP* **0007,** 018 (2000).
39. LHCb Collaboration. *The Large Hadron Collider Beauty Experiment for Precise Measurements of CP Violation and Rare Decays.*<http://lhcb.web.cern.ch/lhcb/> (2007).
40. Blaising, J.-J. *et al. Potential LHC Contributions to Europe's Future Strategy at the High-Energy Frontier.* <http://council-strategygroup.web.cern.ch/council-strategygroup/BB2/contributions/Blaising2.pdf> (2006).
41. FP420 Research and Development Project. FP420 R&D Project <http://www.fp420.com/> (2007).
42. Feng, J. L. & Smith, B. T. Slepton trapping at the CERN Large Hadron Collider and the International Linear Collider. *Phys. Rev. D* **71,** 015004 (2005).
43. Blondel, A. *et al.* Physics opportunities with future proton accelerators at CERN. Preprint at <http://arxiv.org/pdf/hep-ph/0609102> (2006).
44. Strategy Group for European Particle Physics. *CERN Council Strategy Group Home Page* <http://council-strategygroup.web.cern.ch/council-strategygroup/> (2007).
45. Christenson, J. H., Cronin, J. W., Fitch, V. L. & Turley, R. Evidence for the 2π decay of the K_2^0 meson. *Phys. Rev. Lett.* **13,** 138–140 (1964).
46. Kobayashi, M. & Maskawa, T. CP violation in the renormalizable theory of weak interaction. *Prog. Theor. Phys.* **49,** 652–657 (1973).

Author information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence should be addressed to the author (john.ellis@cern.ch).

NATURE|Vol 448|19 July 2007|doi:10.1038/nature06080

The quest for the quark–gluon plasma

Peter Braun-Munzinger[1] & Johanna Stachel[2]

High-energy collisions between heavy nuclei have in the past 20 years provided multiple indications of a deconfined phase of matter that exists at phenomenally high temperatures and pressures. This 'quark–gluon plasma' is thought to have permeated the first microseconds of the Universe. Experiments at the Large Hadron Collider should consolidate the evidence for this exotic medium's existence, and allow its properties to be characterized.

Shortly after the idea of asymptotic freedom — that the interaction between quarks, which is strong at large separations, weakens as the quarks get closer to one another — was introduced by David Gross and Frank Wilczek[1] and David Politzer[2], two groups[3,4] realized independently that it has a fascinating consequence. When temperatures or densities become very high, strongly interacting quarks and gluons become free and transform themselves into a new, deconfined phase of matter, for which the term 'quark–gluon plasma' was coined. We ourselves live at low densities and temperatures, in the normal world of hadronic matter, where quarks and gluons are confined to the size of hadrons. But at its origin, the Universe was a fireball of much higher density and temperature. At times from the electroweak phase transition — some 10 picoseconds after the Big Bang, and lasting for 10 microseconds — it is thought to have taken the form of a quark–gluon plasma. Here, we review the current state of knowledge of this peculiar phase of matter, and outline how the Large Hadron Collider (LHC) should further our understanding of it.

Transition temperature

Various simple estimates lead to a critical temperature for the transition between the familiar, confined hadronic phase of matter and the deconfined, plasma phase of the order of 100 MeV. (In this review, we use the kT unit system, in which all temperatures (T) are multiplied by Boltzmann's constant $k = 8.617 \times 10^{-5}$ eV K^{-1} to express them in more convenient energy units; for reference, a temperature of 100 MeV is somewhat more than 1 trillion kelvin, at 1.16×10^{12} K.) In detailed investigations of hadronic matter, Rolf Hagedorn[5] discovered in the 1960s a limiting temperature for hadronic systems of around the π-meson mass of 140 MeV. It turns out that this temperature is nothing other than the critical temperature for the deconfinement phase transition.

With the advance of solving quantum chromodynamics — the quantum field theory of the strong interaction — on a space-time lattice, more accurate values have become available for the transition temperature. The most readily calculable values are those for a zero net baryon density (that is, no difference between baryon and antibaryon densities). For instance, researchers obtained a temperature[6] of 173 MeV with a systematic error of about 10% for a system involving the two light quark flavours (up and down) and one heavier quark flavour (strange). Very recently, higher values in the vicinity of 190 MeV have been quoted. The reason for the variations is that the lattice energy units have been normalized differently: now, the calculations use quantities that involve heavy bottom quarks, whereas in the past they used the mass of the ρ meson, which contains only the light up and down quarks (see page 270). There is currently a lively debate (see, for example, ref. 7) about the most accurate way to calculate the transition temperature.

Extending lattice quantum chromodynamics into the regime of non-zero net baryon density has met with great technical difficulties. Results[8–10] have become available indicating that the transition temperature drops moderately with increasing density. Going a third of the way from zero net density to the density of atomic nuclei, it drops by 2–3% — not very much. The critical energy density for the phase transition is 0.7 ± 0.2 GeV fm^{-3} (ref. 6). This energy density is about five times that of nuclear matter.

Towards the nuclear fireball

Since the early 1980s, collisions of heavy atomic nuclei at as large energies as possible have been seen as the ideal way to probe these harsh conditions of extremely high temperature and density. To be able to talk about thermodynamic phases, phase transitions, temperatures and so on, the system under consideration must behave like 'matter', not like individual elementary particles or a group of elementary particles. That implies two things. First, the system must consist of a large number of particles (thousands or, better, tens of thousands). Second, it needs to reach local equilibrium, at least approximately, so that variables such as temperature, pressure, energy and entropy density can be defined, and so that thermodynamic relations between those quantities (the equation of state, the speed of sound) can be investigated. This means that the system's lifetime must be significantly larger than the inverse rate of interactions, so that at least a few (order of magnitude five) interactions occur for each constituent, driving the system towards equilibrium.

Collisions of protons (or electrons) produce too few particles to fulfil these conditions. But we know now that collisions between nuclei create enough particles that, if the energy is high enough, they do indeed create a fireball of interacting quarks and gluons above the temperature needed for the phase transition into deconfinement. This fireball quickly expands and cools, until it rehadronizes on passing the deconfinement temperature again. The hugely energetic fireball created in the aftermath of the Big Bang had cooled sufficiently for protons and neutrons (and other confined, but unstable, hadrons) to form after about 10^{-5} s. The fireball created in a nuclear collision in the laboratory contains much less energy, and so is much shorter-lived than that after the Big Bang: after only about 10^{-22} s, the quark–gluon plasma phase of the fireball transforms back to hadronic matter.

Collisions of atomic nuclei have been studied for about 20 years at sufficiently high energies to cross into the deconfined phase. Experimental programmes started simultaneously in late 1986 at the Alternating Gradient Synchrotron at the Brookhaven National Laboratory (BNL) in Upton, New York, and at the Super Proton Synchrotron (SPS)

[1]Gesellschaft für Schwerionenforschung, Planckstr. 1, D 64291, Darmstadt, Germany and Technische Universität Darmstadt. [2]Physikalisches Institut, Universität Heidelberg, Philosophenweg 12, D69120 Heidelberg, Germany.

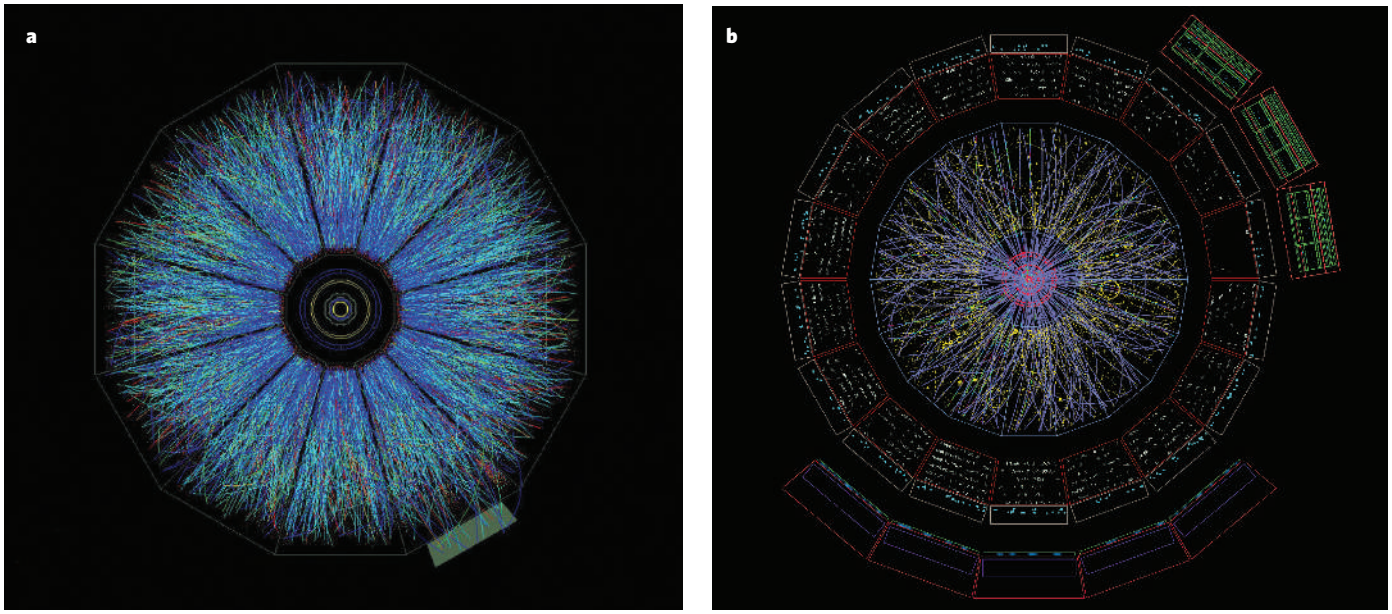


Figure 1 | Fireball remnants. a, Charged particles from a central gold–gold collision at RHIC, recorded by the time projection chamber of the STAR experiment. Colours represent the level of ionization deposited in the detector, with red equating to high values and blue to low values. **b**, A simulation of a central lead–lead collision — just a one-degree slice in polar angle is shown — in the central barrel of the ALICE experiment at the LHC. **Images courtesy of the STAR and ALICE collaborations.**

at CERN. At both facilities, collisions were studied initially with light atomic nuclei (up to silicon and sulphur, with mass numbers of 28 and 32, respectively) and, from the early 1990s, also with heavy nuclei such as gold (mass number 197) and the most abundant isotope of lead (208). For these heavy colliding nuclei, BNL has reached energies in the centre-of-mass system of close to 1,000 GeV, and CERN has reached 3,600 GeV (corresponding to a centre-of-mass energy per colliding nucleon pair, written as $\sqrt{s_{NN}}$, of 4.6 and 17.2 GeV, respectively). At least for the CERN energy regime, enough evidence was gathered to conclude that a new state of matter had been created in these collisions[11,12].

At the same time, a huge next step was being taken. At BNL, a dedicated new accelerator, the Relativistic Heavy-Ion Collider (RHIC), went into operation, servicing four experiments, called BRAHMS, PHENIX, PHOBOS and STAR. At RHIC, heavy nuclei such as gold collide at a relativistic centre-of-mass energy of 40,000 GeV ($\sqrt{s_{NN}} = 200$ GeV). This higher collision energy means a much larger and hotter fireball than had previously been possible (Fig. 1). Data from the first three years of running at RHIC are summarized in refs 13–16, and more recent data can be found in ref. 17.

An even braver new world will come about with the start of operations at the LHC, in which nuclei with masses up to that of lead will be able to collide at a centre-of-mass energy of 1,150 TeV ($\sqrt{s_{NN}} = 5.5$ TeV). This is a huge step in collision energy, about 30 times more than that of RHIC and, at about 0.18 mJ, the first really 'macroscopic' energy to be investigated. The fireball is expected to contain tens of thousands of gluons and quarks, and its temperature should exceed the critical temperature for the deconfinement phase transition several times over. This huge increase in energy should allow the unambiguous identification and characterization of the quark–gluon plasma.

A fireball in chemical equilibrium

As mentioned earlier, one of the crucial questions to be addressed in considering ultra-relativistic collisions between nuclei is the extent to which matter is formed in the fireball. There are two important sets of observations that support the idea of a matter-like fireball. The first concerns the fact that the fireball yields hadrons that are in chemical equilibrium, forming a statistical ensemble. Hadron yields have been studied with high precision in nuclear collisions at the energies used in

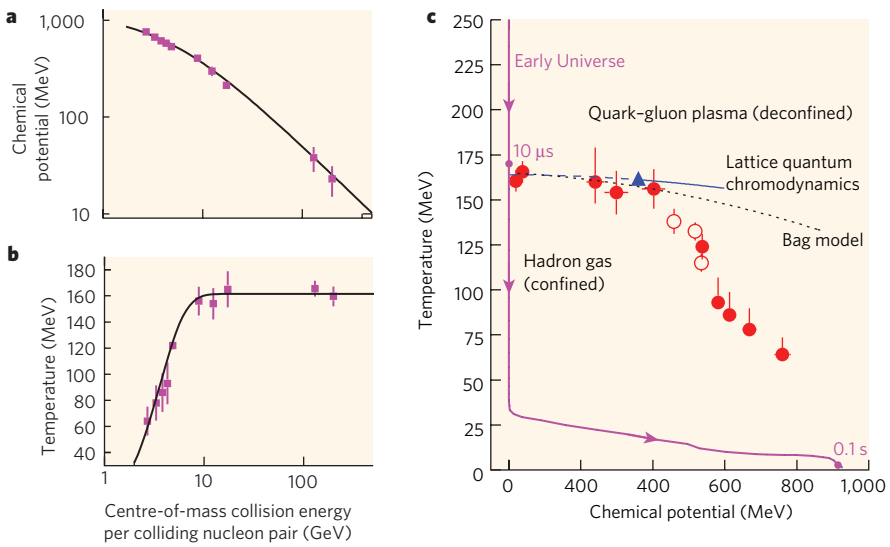


Figure 2 | Equilibrium parameters of the fireball. The energy dependence of chemical potential (**a**) and temperature (**b**), determined from a statistical analysis of hadron yields[18]. The temperature plateau at high collision energies suggests the presence of a phase boundary. Pink squares are results of individual experiments. **c**, The phase diagram of strongly interacting matter: the data points are obtained as in **a** and **b**. The evolution of the early Universe is shown, as are theoretical expectations such as from lattice quantum chromodynamics (blue line) and the bag model (dotted line) for the phase boundary between confined and deconfined matter. Red filled circles are from analysis of midrapidity data. Open circles are from analysis of 4π data. Blue triangle is possible position of a critical endpoint. **Figure reproduced, with permission, from ref. 18.**

the Alternating Gradient Synchrotron, SPS and RHIC. These yields can be described by assuming that all hadrons are formed only when the fireball reaches a specific equilibrium temperature, volume and baryon chemical potential (a measure of the energy change brought about by the addition of one more baryon to the system). Under these conditions, the hadron yields can be characterized in relatively simple terms by the thermodynamic grand-canonical ensemble or, in the special case of small particle numbers, by the canonical ensemble. Such conditions are dubbed the 'chemical freeze-out' scenario, in analogy to the production of bound particles as the early Universe cooled. Detailed analyses of the freeze-out can be found in refs 18 and 19, and a comprehensive review in ref. 20.

Importantly, the energies attained in the SPS and RHIC are also high enough to produce particles containing several strange quarks, including the Ω and $\bar{\Omega}$ baryons. Yields of these baryons agree very well with chemical-equilibrium calculations, and are much higher than in proton–proton collisions. The interpretation is that in heavy-ion collisions, the chemical freeze-out is caused by the quark–gluon plasma and its transition to normal matter, whereas this plasma is absent in collisions between protons.

With increasing centre-of-mass collision energy, the chemical potential decreases smoothly, so new baryons and antibaryons can be created with increasing ease (Fig. 2a). By contrast, although the temperature increases strongly at first, it plateaus rather abruptly near $\sqrt{s_{NN}} = 10$ GeV, at a value slightly higher than 160 MeV (Fig. 2b). This plateau supports Hagedorn's limiting-temperature hypothesis[5], and strongly suggests that a boundary — the phase boundary — is reached at a critical collision energy. Beyond that energy, all additional energy goes into heating the quark–gluon plasma which, in turn, cools again and freezes out at the phase boundary (critical temperature).

If the temperature of the collision fireball is plotted against its chemical potential, with one entry for each energy investigated, a phase diagram can be constructed for the strongly interacting matter contained within it (Fig. 2c). What emerges can be compared to various predictions of the position of the phase boundary taken[8–10] from lattice quantum chromodynamics and[21] from a simple 'bag model' of quarks' confinement into hadrons. For chemical potentials of less than about 400 MeV — corresponding to the critical energy discussed above — the temperatures and chemical potentials determined from the measured hadron yields coincide, within about 10 MeV uncertainty, with the phase boundary as determined from lattice quantum chromodynamics calculations. When the phase boundary is reached, all further points follow it — hadrons cannot be formed in the quark–gluon plasma, only as the plasma rehadronizes.

But could this just be coincidence? What mechanism enforces equilibrium at the phase boundary? Collision rates and the timescales of fireball expansion in the hadronic phase[22] imply that, at the energies used in the SPS and RHIC, equilibrium cannot be established in the hadronic medium. Rather, it is the phase transition between deconfined and confined matter that ensures chemical equilibrium through multi-particle collisions during hadronization. Alternatively, the plateau can be interpreted to arise[23,24] from the filling of phase space during hadronization. In either case, all current interpretations of the observed phenomena relate the chemical variables directly to the phase boundary. This implies that a fundamental parameter of quantum chromodynamics — namely the critical temperature for the deconfinement phase transition — has been determined experimentally to be close to 160 MeV, for small values of chemical potential.

This interpretation will be tested directly by experiments at the LHC. If the plateau phenomenon holds, as is to be expected from the above considerations, then the particle yields measured at LHC energy should, except for an overall volume parameter, agree closely with those measured at the much smaller RHIC energy. That would lend strong support to a phase boundary as the limiting agent.

The observed equilibrium is a strong indication that a matter-like medium is produced in high-energy collisions between nuclei. In collisions among particles such as leptons or nucleons, such equilibrium is not observed, at least not at the energies at which particles containing strange quarks can be produced[20], and hence no medium is formed. We finally note that, as is evident from Fig. 2c, in heavy-ion collisions the chemical freeze-out temperature is not universal but instead varies strongly at large values of the chemical potential. This implies that the properties of the medium change with energy, indicating a transition to a baryon-rich medium at low energies.

The phase transition at low baryon density is probably of the cross-over type[25]. General considerations, as well as results from lattice quantum chromodynamics, suggest the possibility of a first-order phase transition at higher baryon densities with a corresponding critical end-point as sketched in Fig. 2c. Experiments to search for the critical point are planned at the SPS, RHIC and the future Facility for Antiproton and Ion Research (FAIR) at the heavy-ion research centre GSI in Darmstadt, Germany.

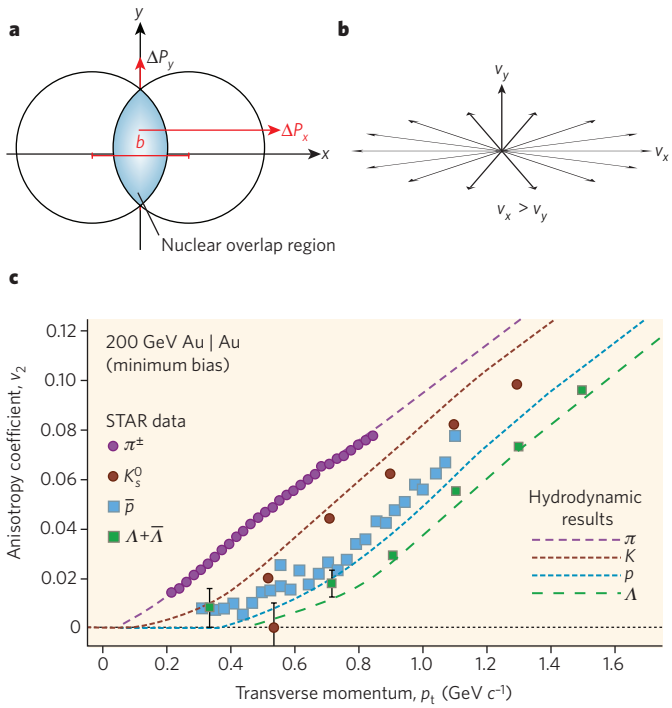


Figure 3 | Geometry of matter during a nuclear collision. a, The nuclear overlap region for semi-central collisions. Early in the collision, the pressure gradient is large in the plane of the collision, *x*. After some time, the large pressure gradient leads to a larger expansion velocity (v_x) in this direction (**b**). The expansion velocity profile in the *x*–*y* plane leads to highly asymmetrical particle emission, with azimuthal anisotropies in momenta perpendicular to the beam (p_t) of various particles. **c**, The distribution of these momenta can be quantified by a Fourier decomposition and parametrized by the second Fourier coefficient $v_2 = \langle\cos 2\varphi\rangle$ — also called the 'elliptical flow'[56–58] — in which the angle φ is measured relative to the direction of impact: higher transverse momenta are recorded for particles emerging in the reaction plane, whereas much lower momenta are observed perpendicular to the reaction plane. As a consequence, the v_2 coefficients are large and show a characteristic p_t dependence. Data for π mesons, *K* mesons, antiprotons ($\bar{p}$) and Λ baryons (with masses mc^2 of about 140, 495, 940 and 1,115 MeV, respectively) agree very well in their mass- and p_t-dependence with predictions[59–61] made with relativistic hydrodynamics and an equation of state determined by weakly interacting quarks and gluons. Although the data are not very sensitive to the particular equation of state used, equations of state based exclusively on hadrons do not lead to a satisfactory description of the data. The data shown are from the STAR experiment at RHIC[62]. **Part c reproduced, with permission, from ref. 62.**

Hydrodynamic expansion and cooling

If matter is formed in the moments after a nuclear collision, hydrodynamic flow effects should be seen owing to the strong pressure gradients present in it. At ultra-relativistic energies, two colliding nuclei are highly

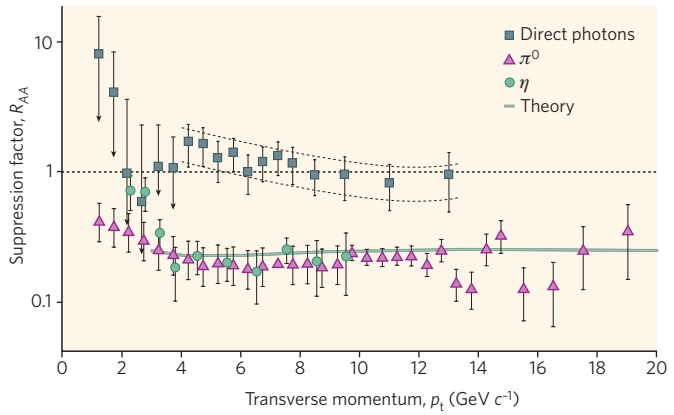


Figure 4 | Preliminary PHENIX results for the suppression factor R_{AA} out to high p_t for π^0 and η mesons. R_{AA} is the ratio of the number of events at different values of p_t for gold–gold collisions normalized to the number of events in proton–proton collisions, scaled by the number of collisions. The suppression of the gold–gold spectrum at high p_t is further evidence for the presence of a hot, dense medium on which jet partons scatter and lose momentum. Bars indicate statistical error. The dotted line at $R_{AA} = 1$ is the expected result when the photon spectrum is unmodified. A theoretical spectrum[63] is also shown, calculated under the assumptions that the precursor parton of the jet loses energy by radiation, and that the medium can be modelled as dense gluon gas. It agrees well with the experimental data. The results for photons produced directly in gold–gold and proton–proton collisions are also shown. These show no suppression; this is consistent with the idea of a gluon gas, as photons do not participate in strong interactions. Figure courtesy of the PHENIX collaboration.

Lorentz-contracted — at the RHIC energy by a factor of 100, at the LHC energy by a factor of 2,700. Consequently, the collision is very quick, lasting around 10^{-25} s. The geometry of the matter immediately after the collision is sketched in Fig. 3a; with increasing impact, b, the overlap zone becomes more and more aspherical in the plane perpendicular to the axis of the colliding beams. It attains an almond-like shape with a typical size in the perpendicular plane determined by the dimensions of the nuclei involved (the diameter of a lead nucleus is about 14 fm), whereas the extension in beam direction cannot be greater than the speed of light multiplied by the collision time — less than 1 fm.

This highly asymmetrical zone evolves by collisions between its constituents (quarks and gluons) until, after a time of about 1 fm c^{-1}, equilibrium is reached and a highly compressed, but still very asymmetrical, fireball is formed. The details of this initial phase are not well understood, but might involve highly coherent configurations of colour fields, generated by a 'colour glass condensate'[26], large fluctuations of which are thought to lead to extremely rapid equilibrium. Irrespective of the details of this highly complex evolution, some ground rules are clear if equilibrium is reached in a short enough time that the shape of the fireball remains essentially unchanged from the initial geometric overlap zone. In this case, the fireball's further evolution should be governed by the laws of relativistic hydrodynamics for a system with very strong pressure gradients, as well as by the equation of state that connects the variables such as volume, temperature and chemical potential that characterize the medium. When the original spatial correlation is transformed into a correlation in momentum (or velocity) space, this implies a very asymmetrical particle emission in the plane perpendicular to the axis of the colliding beams (Fig. 3b). The earlier the equilibrium, and with it the beginning of the hydrodynamic evolution, the larger the anisotropies will be.

What observations are to be expected if the fireball really does expand hydrodynamically, with a unique collective velocity for each fluid cell in the system? The transverse momenta (p_t) of the emitted particles are connected to the fluid velocity via $p_t = m\beta_f\gamma_f$ (with $\gamma_f = 1/\sqrt{1-\beta_f}$), in which m is the mass of the particle, β_f is the relativistic fluid velocity and γ_f is its Lorentz factor, and a characteristic mass-dependent flow pattern arises. The resulting mass ordering in the anisotropy coefficients (v_2) agrees closely with the experimental observations (Fig. 3c). The large anisotropy coefficients confirm the idea that the fireball reaches equilibrium rapidly.

This dramatic and unexpected success is the second strong pillar supporting the idea that ultra-relativistic nuclear collisions produce a collectively expanding medium in thermal equilibrium. We note here that the hydrodynamic calculations with which the experimental data are in such good agreement assume that fluid flow is non-viscous. When added to the hydrodynamic equations, even small viscosity destroys the agreement between data and calculations. Researchers have thus concluded that the matter made in the RHIC fireball is probably close to an ideal fluid[27–29].

If the fireball really does behave hydrodynamically at top RHIC energy, then the elliptical flow data from the LHC should be similar to those from RHIC. The anticipated much greater number of particles produced in each collision could be used to measure flow precisely for many types of particle. That could in turn allow the equation of state of the matter and its transport coefficients (such as viscosity) to be pinned down. Alternatively, as still argued by some authors[30], the flow pattern observed at RHIC is due to a cancellation between unusual initial conditions of the fireball owing to a colour glass condensate on the one hand and an imperfect thermal equilibrium on the other. If that is so, or if viscous effects do play a part, then the LHC data on elliptical flow could reach much larger values. In any case, the very large energy step when going from RHIC to the LHC should lead to important, and urgently needed, new information on the physics of the quark–gluon plasma.

An opaque matter

In collisions of heavy nuclei, hard scattering events — those with high momentum transfer — between the constituent 'partons' (quarks and gluons) liberated are expected to occur just as in collisions between protons. The number of such events, however, will scale with the number of individual proton–proton collisions for a given collision geometry: for head-on collisions of two equal nuclei of mass number A, the number of

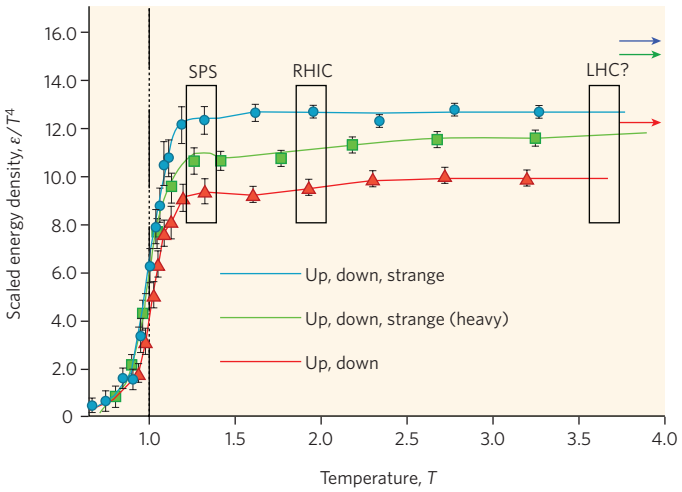


Figure 5 | Scaled energy density ε/T^4 as a function of the temperature calculated in lattice quantum chromodynamics[6]. For an ideal gas, the energy density is proportional to the fourth power of the temperature with the proportionality constant containing the number of degrees of freedom. The strong increase near the critical temperature (T_c, vertical line) indicates that the system is not only heated but that something dramatic happens: it undergoes a phase transition from hadronic matter to quark–gluon plasma with a corresponding large increase in the number of degrees of freedom. Above T_c, the quark–gluon plasma is only heated such that ε/T^4 is constant. The three lines are calculations for two light quark flavours (only up and down; red), three equally light flavours (up, down and strange; blue) and the most realistic case of two light flavours (up and down) and one more massive (strange) flavour (green). Coloured arrows show the expected values of scaled energy density at the Stefan-Boltzman limit. The regions labelled by accelerator facilities indicate maximum initial temperatures reached there. Figure reproduced, with permission, from ref. 65.

events will scale as $A^{4/3}$. Individual collisions between protons are thought to occur independently of each other, and their number can be computed from the distributions of the nuclear densities, the nuclear overlap for a given impact and the inelastic proton–proton cross-section.

Collisions of nuclei differ from collisions between protons in that the hard scattered partons may traverse the quark–gluon plasma before or during their hadronization into a jet. Jets are characteristic of collisions between protons in which two constituent partons scatter and recede from each other with a significant fraction of the initial beam momentum. In the plane transverse to the beams, the momenta are large and opposite in direction. The two scattered partons hadronize mainly into mesons that are emitted in a cone — the jet — around the direction of parton momentum. It was realized very early[31] that the quark–gluon plasma could modify jets resulting from collisions between nuclei. Calculations showed that a parton traversing a hot and dense medium consisting of other partons — that is, a quark–gluon plasma — should lose substantially more energy than one traversing cold nuclear matter[32–34]. This prediction appears to be borne out by data from all four experiments at RHIC.

A jet is much more difficult to see in a heavy-ion collision than after a collision between protons. The reason is the sheer number of particles produced: a single central (head-on) gold–gold collision generates about 5,000 charged particles, and unless the jet has very high (transverse) momentum, it will not stand out in the crowd. But the presence of jets will affect the overall transverse momentum distribution. At low transverse momenta, the spectrum in a heavy-ion collision is complex, as it is a superposition of hydrodynamic expansion effects and random thermal motion. Nevertheless, for particles of a particular species with transverse momenta that are significantly larger than their mass, the resulting spectrum is nearly exponential. The contribution of jets with high transverse momentum leads to a distinct power-law behaviour typically visible for values of transverse momentum of a few GeV or more.

To judge a possible modification of the shape of the spectrum in a high-energy nuclear collision, the transverse-momentum distribution of π mesons produced in central gold–gold collisions at RHIC can be compared with that measured in proton–proton collisions. To quantify this comparison, the ratio of the gold–gold-collision spectrum to the proton–proton-collision spectrum is scaled to the total number of inelastic collisions in the nuclear case, providing the suppression factor R_{AA}. For larger transverse momenta, this factor settles at about 0.2 (Fig. 4); that is, the production of high-momentum π mesons is suppressed by a factor of five in gold–gold collisions.

What is the origin of this suppression? The transverse-momentum spectrum for collisions between protons agrees well[35] with theoretical calculations that use next-to-leading-order quantum chromodynamic perturbation theory. When the spectra of deuteron–gold collisions of varying centrality are compared with the proton–proton spectrum, R_{AA} is 1 or larger (for more central collisions, values larger than 1 are even expected — a phenomenon known as the Cronin effect, caused by the scattering of partons before the hard collision). For peripheral gold–gold collisions, the values of R_{AA} also correspond well to the expectation from collisions between protons. The clear implication is that something special and new happens in central gold–gold collisions: the precursor parton of the jet produced must lose a lot of energy, causing the transverse-momentum spectrum of the mesons in the jet to fall off steeply.

Several researchers have shown that only calculations including large energy loss in the medium can account for these data. The clear implication is that the medium present in the collision fireball is hot and dense, and when partons pass through it, they lose energy. Both radiation of gluons and elastic scattering seem to be important here. In deuteron–gold collisions, by contrast, the jet sees at most cold nuclear matter (or a vacuum), and does not seem to be perturbed.

Calculating the energy loss of a fast parton in a quantum chromodynamic liquid, as suggested by the data discussed in the previous section, is beyond the current theoretical state-of-the-art. To gain insight into the underlying physics of energy loss, it is helpful to resort to another aspect of the medium: that it contains many gluons. Indeed, the RHIC data on parton energy loss are well explained by modelling the medium formed by the collision as an ultra-dense gluon gas with a density of the number of gluons (N_g) per rapidity interval of $dN_g/dy = 1{,}100$. Here, the rapidity y is a logarithmic measure of the gluon's longitudinal velocity, v. With the simple assumption that $v = z/t$ (z is the longitudinal space coordinate), Bjorken[36] showed how to map rapidity densities to spatial densities. The spatial gluon density in turn is linked directly to entropy density. Using relations from statistical mechanics for a relativistic gas of bosons (and fermions if quarks are included), the temperature and energy density can be obtained from these gluon densities. The high gluon densities needed to reproduce the observed gold–gold R_{AA} correspond to an initial temperature of about twice the critical temperature for the formation of a quark–gluon plasma. The initial energy densities of 14–20 GeV fm^{-3} are

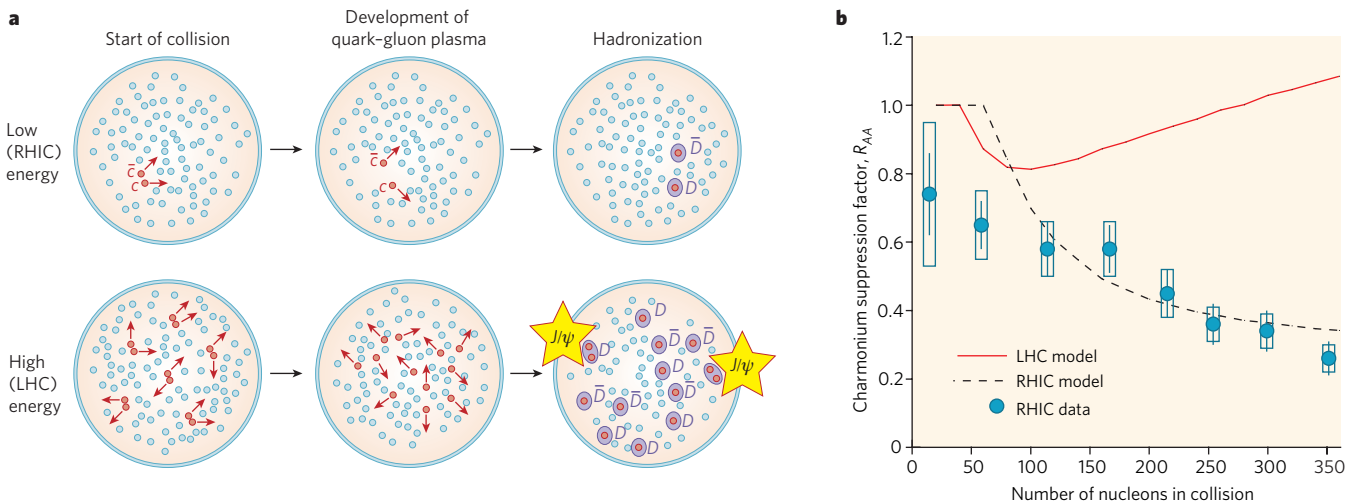


Figure 6 | Charmonium suppression. a, At low energies, the quark–gluon plasma screens interaction between the only pair of charm quark and antiquark produced (red dots) and any other two quarks (up, down, strange) will find themselves paired with the charm quark/antiquark in *D* mesons at hadronization (purple circles). At high energies, by contrast, many charm–anticharm pairs are produced in every collision and at hadronization, charm and anticharm quarks from different original pairs may combine to form a charmonium J/Ψ particle. Grey dots indicate light partons produced in the collision. **b**, Theory and experiment compared quantitatively. Model predictions[55] for the charmonium suppression factor agree well with recent RHIC data from the PHENIX collaboration[66]. Owing to the increased level of statistical recombination expected, enhancement rather than suppression is predicted for LHC conditions. What the experiments deliver will be a further crucial test of theories of the quark–gluon plasma. Part b reproduced, with permission, from ref. 55.

Box 1 | ALICE: the LHC's dedicated plasma hunter

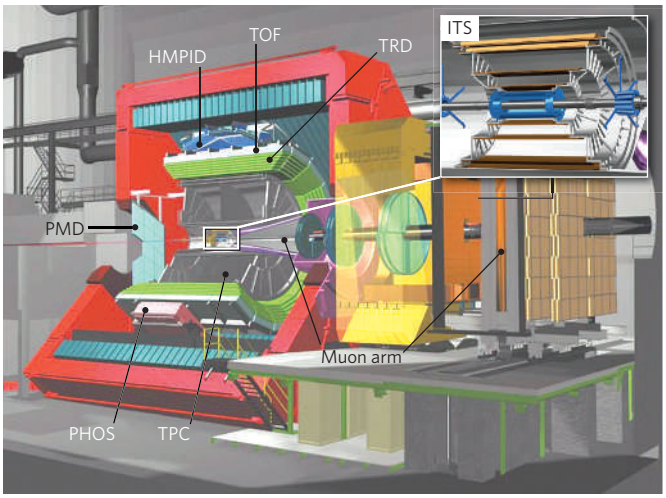
LHC's ALICE detector is dedicated to investigating nucleus–nucleus collisions at an energy of 2.8 TeV per nucleon in each of the colliding nuclei. The main part of the apparatus is housed in the world's largest solenoidal magnet, which generates a field strength of 0.5 T within a volume of 1,600 m^3.

Various detectors arranged in cylindrical shells around the interaction point (see figure) are designed to determine the identity and precise trajectory of the more than 10,000 charged particles propelled by a lead–lead nuclear collision into the active volume of the apparatus. The innermost detector is the inner tracking system (ITS), which consists of six layers of silicon detectors surrounding the 1-mm-thick beam pipe that encloses the ultra-high vacuum of the accelerator. These detectors are capable of high-precision tracking (resolution around 20 μm) so as to determine the decay vertex of short-lived particles carrying strange, charm, or bottom quantum numbers that typically decay within a few millimetres to centimetres of the primary interaction point.

The ITS is contained within, and mounted on, the cylindrical barrel of the time projection chamber (TPC). This is ALICE's major tracking device: it is the largest of its kind worldwide, with some 560,000 readout channels, and provides essentially continuous, three-dimensional tracking of charged particles between radii of 80 cm and 250 cm from the central interaction point.

Outside the TPC are two very large (with areas of around 150 m^2) particle identification detectors: the transition radiation detector (TRD), with more than 1 million channels and an on-board computer farm of a quarter-million central processing units for the triggering and identifying of electrons; and surrounding this, the time-of-flight (TOF) detector which can record the transit time between the interaction point and the detector surface at a resolution of better than 100 picoseconds.

The central barrel of ALICE is completed by dedicated detectors to measure photons (PHOS) and their distribution in the forward direction (PMD) and to identify high-momentum hadrons (HMPID), and by



further detectors to determine the position and time of the primary interaction point. Separated from the main detector in the forward direction of one of the accelerator beams, and behind a conical absorber that projects into the central barrel, is a muon detector with its own large dipole magnet. Because muons do not undergo strong reactions, and because those at relevant energies do not emit *bremsstrahlung*, they penetrate the absorber practically unscathed — unlike hadrons. Their momenta are then measured in tracking stations before, inside, and after the dipole magnet.

The huge scale and cutting-edge engineering of the ALICE detector should allow it to make a decisive contribution to understanding the properties of the medium that will be created in the LHC's high-energy collisions between nuclei. Image courtesy of the ALICE collaboration.

also well in line with the initial conditions required by the hydrodynamic models introduced earlier. Both the temperature and the energy density are well above the critical conditions calculated with lattice quantum chromodynamics (Fig. 5).

Another important cross-check for the gluon-gas interpretation is to compare the transverse-momentum spectrum of photons produced in hard initial parton scattering in gold–gold and in proton–proton collisions. For most values of transverse momentum, the corresponding R_{AA} factors are consistent with unity (Fig. 4). This is perfectly in line with an unmodified distribution as produced in initial hard scattering: the photon does not participate in the strong force, and so would traverse a gluon-dominated fireball without further interaction.

Similar analyses have been done for several different hadronic species. Generally, within the errors, all mesons behave just like π mesons (the data points for η mesons are shown in Fig. 4). In an intermediate range of 2–6 GeV, the suppression of baryons is significantly weaker, but for high transverse momentum it joins that for π mesons.

Another characteristic of hadron jets is the back-to-back correlation of two of them in the plane perpendicular to the colliding beams. Even without reconstructing the jet (the difficulty of doing this in a heavy-ion collision has been mentioned), such an analysis can be done by selecting just the particles that have the highest momentum in an event. Both the PHENIX and STAR experiments have undertaken such a 'leading particle' analysis, picking just one high-transverse-momentum trigger particle (say, in the range 4–6 GeV) and then checking for the distribution in azimuthal angle of all other 'associated' particles in a given range of transverse momenta. Proton–proton experiments produce two peaks, one at 0° and one at 180°. This is expected, as the associated particle could be a fragment of the same jet (the 0° peak), or a leading particle of the second jet diametrically opposite (the 180° peak).

In gold–gold collisions, a dramatic change is seen[37,38] when the transverse momentum of the associated particle is varied. At very high momenta, the two expected peaks are present. For lower particle momenta (2–4 GeV), however, the peak at 180° broadens to the point that it is barely visible. For even lower momenta, it develops into a dip with two pronounced peaks, one on either side about 1 radian apart. A very similar observation has been made at the top SPS energy[39,40]. The interpretation of this feature of the data is still hotly debated, but an interesting suggestion has been made[41,42]: the dip with the two satellite peaks could indicate the presence of a shock wave in the form of a Mach cone caused by a supersonic parton traversing the quark–gluon plasma. If borne out, this could lead to the determination of the speed of sound in the plasma.

The modelling of the parton energy loss in the quark–gluon plasma is still somewhat schematic, and leaves open a range of theoretical possibilities and ways of implementation. To sharpen the interpretation, it would be good to have individual measurements of the properties of jets stemming from all different quark flavours, as well as from gluons. For heavy quarks (charm and bottom), an important step has been made by the PHENIX and STAR collaborations at RHIC. Both experiments have measured transverse-momentum spectra of electrons stemming from decays of D and B mesons (each of which contain a charm or bottom quark) into electrons plus anything else. The R_{AA} values have been determined for these spectra[43,44], too, and have been surprising: they are very close to the values determined for mesons that involve only up, down and strange quarks. Theoretically, energy loss by radiation should be much lower for the charm and bottom quarks than for the up, down and strange quarks and gluons owing to their much larger masses. Since then, it has been realized that energy loss by scattering is probably of comparable importance to energy loss by radiation[45,46], which improves the quantitative situation somewhat, but doesn't resolve the puzzle.

With the start of experiments at the LHC, matters will change dramatically: because of the much higher beam energy, jet production will be enhanced by many orders of magnitude compared with the situation at RHIC energy. Estimates[47] based on solving quantum chromodynamics by perturbation theory imply, for example, an enhancement of

more than four orders of magnitude at $p_t = 100$ GeV. Thus a whole new range of transverse-momentum values between 20 and 250 GeV will become accessible. Such measurements can then be used to discriminate between the various theoretical scenarios competing to describe parton energy loss, either by determination of R_{AA} or, uniquely for the LHC, by direct reconstruction of the jet in a collision between nuclei. This should allow jet probes to be developed into quantitative tools to determine the parton density of the matter formed.

Charmonia as harbingers of deconfinement?

The particles collectively known as charmonia — bound states of heavy charm quarks and antiquarks — have a special role in research into the quark–gluon plasma. In 1986, Satz and Matsui[48] realized that the high density of gluons in a quark–gluon plasma should destroy charmonium systems, in a process analogous to Debye screening of the electromagnetic field in a plasma through the presence of electric charges. The suppression of charmonia (compared with their production in the absence of a quark–gluon plasma) was thus proposed as a 'smoking gun' signature for plasma formation in nuclear collisions at high energy. Measurements at the SPS[49] did indeed provide evidence for such suppression in central collisions between heavy nuclei. No suppression was found in grazing collisions or in collisions between light nuclei, in which a plasma is not expected to form. But absorption of charmonium in the nuclear medium, as well as its break-up by hadrons produced in the collision, is also a mechanism that could lead to charmonium suppression even in the absence of plasma formation[50], and the interpretation of the SPS data remains inconclusive.

This situation took an interesting turn in 2000, when researchers realized that the large number of charm-quark pairs produced in nuclear collisions at collider energies could lead to new ways to produce charmonium, either through statistical production at the phase boundary[51,52], or through coalescence of charm quarks in the plasma[53]. At low energy, the mean number of charm-quark pairs produced in a collision is much fewer than 1, implying that charmonium is formed, if at all, always from charm quarks of this one pair. Because the number of charm quarks in a collision at LHC energy is expected to reach about 200, charm quarks from different pairs can combine to form charmonium (Fig. 6a). This works effectively only if a charm quark can travel a substantial distance in the plasma to 'meet' its prospective partner. Under these conditions, charmonium production scales quadratically with the number of charm-quark pairs, so enhancement, rather than strong suppression, is predicted for LHC energy[54,55] (Fig. 6b). If observed, this would be a spectacular fingerprint of a high-energy quark–gluon plasma, in which charm quarks are effectively deconfined. Again, as in most other cases, the data from the LHC will be decisive in settling the issue.

Looking forward

The data from the SPS and particularly the RHIC accelerator have taught us that central nuclear collisions at high energy produce a medium made up of partonic matter in equilibrium and possessing collective properties. The medium flows much like an ideal liquid and is dense enough to dissipate most of the energy of a 20 GeV parton. Future experiments at RHIC will further elucidate some of the aspects discussed above, in particular in the heavy-quark sector. With their 30 times higher energy, lead–lead collisions at LHC in the specially designed ALICE detector (Box 1) will produce this new state of matter at unprecedented energy densities and temperatures and over very large volumes compared with the size of the largest stable nuclei. With the planned experiments, the LHC heavy-ion community looks forward with anticipation to elucidating the properties of such partonic fireballs. The prize is unravelling the mystery of the matter that formed a fraction of a nanosecond after the Big Bang, and disappeared just 10 microseconds later. ■

1. Gross, D. J. & Wilczek, F. Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.* **30,** 1343–1346 (1973).
2. Politzer, H. J. Reliable perturbative results for strong interactions? *Phys. Rev. Lett.* **30,** 1346–1349 (1973).
3. Cabibbo, N. & Parisi, G. Exponential hadronic spectrum and quark liberation. *Phys. Lett. B* **59,** 67–69 (1975).
4. Collins, J. C. & Perry, M. J. Superdense matter: neutrons or asymptotically free quarks? *Phys. Rev. Lett.* **34,** 1353–1356 (1975).
5. Hagedorn, R. Statistical thermodynamics of strong interactions at high energies. *Nuovo Cimento Suppl.* **3,** 147–186 (1965).
6. Karsch, F., Laermann, E. & Peikert, A. Quark mass and flavour dependence of the QCD phase transition. *Nucl. Phys. B* **605,** 579–599 (2001).
7. Aoki, A., Fodor, Z., Katz, S. D. & Szabo, K. K. The QCD transition temperature: results with physical masses in the continuum limit. *Phys. Lett B.* **643,** 46–54 (2006).
8. Fodor, Z. & Katz, S. Critical point of QCD at finite T and μ, lattice results for physical quark masses. *J. High Energy Phys.* **4,** 050 (2004).
9. Allton, C. R. *et al.* The QCD thermal phase transition in the presence of a small chemical potential. Preprint at <http://arxiv.org/abs/hep-lat/0204010> (2002).
10. Ejiri, S. *et al.* Study of QCD thermodynamics at finite density by Taylor expansion. Preprint at <http://arxiv.org/abs/hep-lat/0312006> (2003).
11. CERN. New state of matter created at CERN. <http://press.web.cern.ch/press/PressReleases/Releases2000/PR01.00EQuarkGluonMatter.html> (2000).
12. Heinz, U. & Jacob, M. Evidence for a new state of matter: an assessment of the results from the CERN lead beam programme. Preprint at <http://arxiv.org/abs/nucl-th/0002042> (2000).
13. Arsene, I. *et al.* Quark–gluon plasma and color glass condensate at RHIC? The perspective from the BRAHMS experiment. *Nucl. Phys. A* **757,** 1–27 (2005).
14. Back, B. B. *et al.* The PHOBOS perspective on discoveries at RHIC. *Nucl. Phys. A* **757,** 28–101 (2005).
15. Adams, J. *et al.* Experimental and theoretical challenges in the search for the quark–gluon plasma: the STAR Collaboration's critical assessment of the evidence from RHIC collisions. *Nucl. Phys. A* **757,** 102–183 (2005).
16. Adcox, K. *et al.* Formation of dense partonic matter in relativistic nucleus–nucleus collisions at RHIC: experimental evaluation by the PHENIX Collaboration. *Nucl. Phys. A* **757,** 184–283 (2005).
17. Csörgö, T., Dávid, G., Lévai. P. & Papp, G. (eds) Quark matter 2005 — proceedings of the 18th international conference on ultra-relativistic nucleus–nucleus collisions. *Nucl. Phys. A* **774,** 1–968 (2006).
18. Andronic, A., Braun-Munzinger, P. & Stachel, J. Hadron production in central nucleus–nucleus collisions at chemical freeze-out. *Nucl. Phys. A* **772,** 167–199 (2006).
19. Becattini, B., Gadzicki, M., Keranen, A., Manninen, J. & Stock, R. Chemical equilibrium study in nucleus–nucleus collisions at relativistic energies. *Phys. Rev. C* **69,** 024905 (2004).
20. Braun-Munzinger, P., Redlich, K. & Stachel, J. in *Quark–Gluon Plasma 3* (eds Hwa, R. C. & Wang, X. N.), 491–599 (World Scientific, Singapore, 2004).
21. Braun-Munzinger, P. & Stachel, J. Probing the phase boundary between hadronic matter and the quark-gluon plasma in relativistic heavy-ion collisions. *Nucl. Phys. A* **606,** 320–328 (1996).
22. Braun-Munzinger, P., Stachel, J. & Wetterich, C. Chemical freeze-out and the QCD phase transition temperature. *Phys. Lett. B.* **596,** 61–69 (2004).
23. Stock, R. The parton to hadron phase transition observed in Pb+Pb collisions at 158 GeV per nucleon. *Phys. Lett. B* **456,** 277–282 (1999).
24. Heinz, U. Hadronic observables: theoretical highlights. *Nucl. Phys. A* **638,** c357–364 (1998).
25. Aoki, Y., Fodor, Z., Katz, S. D. & Szabo, K. K. The order of the quantum chromodynamics transition predicted by the standard model of particle physics. *Nature* **443,** 675–678 (2006).
26. Gyulassy, M. & McLerran, L. New forms of QCD matter discovered at RHIC. *Nucl. Phys. A* **750,** 30–63 (2005).
27. Baier, R., Romatschke, P. Causal viscous hydrodynamics for central heavy-ion collisions. Preprint at <http: //arxiv.org/nucl-th/0610108> (2006).
28. Romatschke, P. Causal viscous hydrodynamics for central heavy-ion collisions II: meson spectra and HBT radii. Preprint at <http://arxiv.org/nucl-th/0701032> (2007).
29. Heinz, U., Song, H. & Chaudhuri, A. K. Dissipative hydrodynamics for viscous relativistic fluids. *Phys. Rev. C* **73,** 034904 (2006).
30. Bhalerao, R. S., Blaizot, J. P., Borghini, N. & Ollitrault, J. Y. Elliptic flow and incomplete equilibration at RHIC. *Phys. Lett. B* **627,** 49–54 (2005).
31. Bjorken, J. D. Fermilab-PUB-82-59-THY (1982) and erratum (unpublished).
32. Wang, X. N. & Gyulassy, M. Gluon shadowing and jet quenching in *A+A* collisions at $\sqrt{s} = 200A$ GeV. *Phys. Rev. Lett.* **68,** 1480–1483 (1992).
33. Baier, H., Dokshitzer, Y. L., Mueller, A. H., Peigne, S. & Schiff, D. Radiative energy loss of high energy quarks and gluons in a finite-volume quark-gluon plasma. *Nucl. Phys. B* **483,** 291–320 (1997).
34. Baier, H., Dokshitzer, Y. L., Mueller, A. H., Peigne, S. & Schiff, D. Radiative energy loss and $p_\perp$-broadening of high energy partons in nuclei. *Nucl. Phys. B* **484,** 265–282 (1997).
35. d'Enterria, D. Quantum chromo (many-body) dynamics probed in the hard sector at RHIC. Preprint at <http://arxiv.org/abs/nucl-ex/0406012> (2004).
36. Bjorken, J. D. Highly relativistic nucleus–nucleus collisions: the central rapidity region. *Phys. Rev. D* **27,** 140–151 (1983).
37. Adler, S. S. *et al.* (PHENIX collaboration). A detailed study of high-p_T neutral pion suppression and azimuthal anisotropy in Au+Au Collisions at $\sqrt{s_{NN}} = 200$ GeV. Preprint at <http://arxiv.org/abs/nucl-ex/0611007> (2006).
38. Mischke, A. (for the STAR collaboration). High-p_T hadron production and triggered particle correlations. Preprint at <http://arxiv.org/abs/nucl-ex/0605031> (2006).
39. Adamova, D. *et al.* (CERES collaboration). Semihard scattering unraveled from collective dynamics by two-pion azimuthal correlations in 158*A* GeV/*c* Pb+Au collisions. *Phys. Rev. Lett.* **92,** 032301 (2004).
40. Ploskon, M. (for the CERES collaboration). Two particle azimuthal correlations at high transverse momentum in Pb–Au at 158 A GeV/c. Preprint at <http://arxiv.org/abs/nucl-ex/0701023nucl-ex/0701023> (2007).
41. Stöcker, H. Collective flow signals the quark–gluon plasma. *Nucl. Phys. A* **750,** 121–147 (2005).

42. Casalderrey-Solana, J., Shuryak, E. & Teaney, D. Conical flow induced by quenched QCD jets. Preprint at <http://arxiv.org/abs/hep-ph/0411315> (2004).
43. Adare, A. *et al.* (the PHENIX collaboration). Energy loss and flow of heavy quarks in Au+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV. Preprint at <http://arxiv.org/abs/nucl-ex/0611018> (2006).
44. Zhang, H. *et al.* Heavy flavour production at STAR. *J. Phys. G* **32,** S29–S34 (2006).
45. Zapp, K., Ingelman, G., Rathmans, J. & Stachel, J. Jet quenching from soft QCD scattering in the quark–gluon plasma. *Phys. Lett. B.* **637,** 179–184 (2006).
46. Adil, A., Gyulassy, M., Horowitz, W. A. & Wicks, S. Collisional energy loss of non asymptotic jets in a QGP. Preprint at <http://arxiv.org/abs/nucl-th/0606010> (2006).
47. Vitev, I. Contribution to CERN yellow report on hard probes in heavy ion collisions at the LHC. Preprint at <http://arxiv.org/abs/hep-ph/0310274> (2003).
48. Satz, H. & Matsui, T. J/ψ suppression by quark-gluon plasma formation. *Phys. Lett. B* **178,** 416–422 (1986).
49. Abreu. M. C. *et al.* Transverse momentum distributions of $J/\psi, \psi'$. Drell–Yan and continuum dimuons produced in Pb–Pb interactions at the SPS. *Phys. Lett. B* **499,** 85–96 (2001).
50. Capella, A., Kaidalov, A. B. & Sousa, D. Why is the J/ψ suppression enhanced at large transverse energy? *Phys. Rev. C* **65,** 054908 (2002).
51. Braun-Munzinger, P. & Stachel, J. (Non)thermal aspects of charmonium production and a new look at J/ψ suppression *Phys. Lett. B* **490,** 196–202 (2000).
52. Braun-Munzinger, P. & Stachel, J. On charm production near the phase boundary. *Nucl. Phys. A* **690,** 119–126 (2001).
53. Thews, R. L., Schroedter, M. & Rafelski, J. Enhanced J/ψ production in deconfined quark matter. *Phys. Rev. C* **63,** 054905 (2001).
54. Andronic, A., Braun-Munzinger, P., Redlich, K. & Stachel, J. *Nucl. Phys. A.* Preprint at <http://arxiv.org/abs/nucl-th/0611023> (2006).
55. Andronic, A., Braun-Munzinger, P., Redlich, K. & Stachel, J. *Phys. Lett. B.* Preprint at <http://arxiv.org/abs/nucl-th/0701079> (2007).
56. Barrette, J. *et al.* Observation of anisotropic event shapes and transverse flow in Au–Au collisions at AGS energy. *Phys. Rev. Lett.* **73,** 2532–2536 (1994).
57. Voloshin, S. & Zhang, Y. C. Flow study in relativistic nuclear collisions by Fourier expansion of azimuthal particle distributions. *Z. Phys. C* **70,** 665–671 (2007).
58. Poskanzer, A. M & Voloshin, S. A. Methods for analyzing anisotropic flow in relativistic nuclear collisions. *Phys. Rev. C* **58,** 1671–1678 (1998).
59. Huovinen, P. *et al.* Radial and elliptic flow at RHIC: further predictions. *Phys. Lett. B* **503,** 58–64 (2001).
60. Teaney, D., Lauret, J. & Shuryak, E. A hydrodynamic description of heavy ion collisions at the SPS and RHIC. Preprint at <http://arxiv.org/abs/nucl-th/0110037> (2001).
61. Kolb, P. & Heinz, U. in *Quark–Gluon Plasma 3* (eds Hwa, R. C. & Wang, X. N.) 634–714 (World Scientific, Singapore, 2004).
62. Adams, J. *et al.* (the STAR collaboration). Azimuthal anisotropy in Au+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV. *Phys. Rev.* **C72,** 014904 (2005).
63. Vitev, I. Jet quenching in relativistic heavy ion collisions *J. Phys. G.* Preprint at <http://arxiv.org/abs/hep-ph/0503221> (2005).
64. Akiba, Y. Probing the properties of dense partonic matter at RHIC. *Nucl. Phys. A* **774,** 403–408 (2006).
65. Kolb, P. & Heinz, U. in *Quark–Gluon Plasma 3* (eds Hwa, R. C. & Wang, X. N.) 1–59 (World Scientific, Singapore, 2004).
66. Adare, A. *et al.* (PHENIX collaboration). J/ψ production vs centrality, transverse momentum, and rapidity in Au+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV. Preprint at <http://arxiv.org/abs/nucl-ex/0611020nucl-ex/0611020> (2006).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing interests. Correspondence should be addressed to J.S. (stachel@physi.uni-heidelberg.de).

NATURE|Vol 448|19 July 2007|doi:10.1038/nature06081

The God particle *et al.*

Leon Lederman

The territory of the Large Hadron Collider might be populated not just by the Higgs particle but also by all manner of other exotic apparitions.

The birth of particle physics — that is, high-energy physics — can be dated to about 1950, offspring of the marriage of nuclear physics and the study of cosmic rays. It exploited techniques and technology from both disciplines, and its objective was to identify the primordial particles of nature — those from which all matter is made — and codify the laws of physics that oversee their properties and social behaviours.

Progress in high-energy physics has always been mortgaged to the requirements of the ever more powerful particle accelerators required to reveal the inner life of the particle 'zoo'. Around 1950, the world energy record was held by a synchrocyclotron accelerator that accelerated protons to 400 megaelectronvolts (MeV) around a circular path. That was enough to shatter atomic nuclei and produce copious quantities of pions and muons. These particles, first discovered in investigations of cosmic rays, were indicators of a vast complexity to come.

The record-breaking 'atom-smasher' of 1950 was constructed by the physics department of Columbia University on the Nevis estate, about 30 miles north of the university, bordering the Hudson River. At its dedication, by the then University president, Dwight Eisenhower, a series of relays brought the Nevis synchrocyclotron to life, as attested by a Geiger counter emitting an amplified series of clicks. My job, as a new graduate student intent on using the accelerator for my PhD research, was to stand by with a radioactive source in case the machine failed. It did, of course, and, as I had misplaced the source, the dawn of this new era in particle physics was delayed, in the ears of the assembled company, by five very embarrassing minutes.

The Nevis machine soon left this small setback behind. This machine, and others that followed all over the world, ensured that by 1995 the cutting-edge energy domain had climbed by a factor of more than 2,000 over those early days. The probe of choice was protons at an energy of 900 gigaelectronvolts (GeV) in head-on collision with antiprotons of the same energy. These collisions, which occurred at a rate of almost 10^6 per second, took place in the 6.3-km-circumference ring of the Tevatron accelerator, at Fermilab, in Batavia, Illinois. The principle of conservation of momentum tells us that a head-on collision is much more violent than is aiming one beam at a stationary target, as the Nevis machine had done. It is the difference between a large speeding truck colliding with a ping-pong ball and two equally huge trucks involved in a full-on collision. In the first case, nothing much happens to the truck, and the ping-pong ball recoils rapidly, none the worse for wear. In the second case, bumpers, mirrors, radios and steering wheels fly off in all directions. Picking through the debris left by such an impact gives us a good grasp of how the truck's interior was put together.

In 1979, as the new director of Fermilab, I made the decision that protons should smash into antiprotons in the new accelerator. The Tevatron's success was crowned in 1995 with its discovery of the last and heaviest of the expected fundamental matter particles: the top quark. The ultimate product of the increasingly savage collisions at Fermilab and elsewhere in the years between 1950 and 1995 was the seemingly complete, self-contained and self-consistent table of nature's fundamental particles — the 'standard model' (see page 270).

With this table now seemingly replete, why is there still hunger for further discovery? Why is the Tevatron now running 24 hours a day, 7 days a week at huge, historic collision rates for fear of what its soon-to-be rival, the Large Hadron Collider (LHC), might find? There is a palpable sense of expectation in the control rooms of the Tevatron; in the construction activities of the physicists from all over the world participating in the LHC project; in the coming together of the massive detectors that will provide the eyes of the LHC; and, most especially, in the quiet rooms populated by theoretical physicists.

A new frontier

By 2009, it is reasonable to expect that the LHC will have claimed the crown of king of the accelerators. When the LHC is completed, the frontier of particle physics will be at a total collision energy of 14 teraelectronvolts (TeV), far beyond the energies reached by the Tevatron. At this frontier, the past decade of research in high-energy physics and in experimental astrophysics tells us that the known, explored world of the standard model and the summed achievements of the past half-century will be behind us. Defects in our theoretical construct — founded on the twin pillars of quantum field theory and Albert Einstein's general theory of relativity — that have become ever more apparent, but that could until now be ignored, will have to be confronted. We are reaching what the medieval map-maker would have denoted *terra incognita*.

By far the most popular expected denizen of these unknown lands is the Higgs particle, which was postulated to tidy up the glitch in the standard model known as electroweak symmetry breaking. Ever since Einstein published his special theory of relativity in 1905, theorists have had great respect for symmetry. At the heart of special relativity is the idea of Lorentz symmetry: that all laws of physics should be the same for all observers moving with constant relative velocities. The equation $E = mc^2$ is a direct consequence of this symmetry. Today, you cannot visit a high-energy physics laboratory without stumbling over symmetry on your way in. As in art and architecture, symmetry in this sense is an aesthetic concept: we believe that nature is best described in equations that are as simple, beautiful, compact and universal as possible. According to this way of thinking, the W and Z particles, which carry the weak nuclear force, and the photon, carrier of the electromagnetic force, should combine to show electroweak symmetry, and all should have zero mass.

> "We are reaching what the medieval map-maker would have denoted *terra incognita*."

Unless the aesthetes are fundamentally wrong, therefore, the fact that electroweak symmetry isn't perfect — because the W and Z particles are heavy — means that something is acting to break the symmetry. This something will give mass not only to the W and Z particles but also to all other particles except those few (such as photons) that can escape its clutches. The effect can be compared to running swiftly on hard ground versus knee-high through oil. In oil, your motion is slower, as if your mass had increased. The Higgs particle is that oil, and a Higgs 'field' is spread across the entire Universe.

In 1993, I co-authored a book on the history and status of high-energy physics. Then, as now, this mysterious Higgs field haunted us. As well

as explaining why particles had mass, the Higgs field allowed theorists to calculate reactions that, lacking such a speculative field, yielded nonsensical values. The beauty of the Higgs idea stimulated us to name the book *The God Particle*. "Besides," as my editor explained with an eye on the sales figures, "no one has ever heard of Higgs."

Ever more precise data emerging from the Tevatron indicate that the Higgs particle itself is not very heavy. It should be relatively easy to produce at the huge energy of the LHC. If so, questions abound. Is the Higgs alone, or is there a whole family of Higgs-type particles? Does the Higgs really give mass to everything: not just to the *W* and *Z* particles but to quarks and charged leptons as well? That really would be the key to a unified theory embracing the gamut of particle physics. But even just knowing the role of the Higgs in breaking electroweak symmetry will allow us to gain understanding of that symmetry and add consistency to quantum field theory. For the sake of its own consistency, the standard model needs something like a Higgs.

The Higgs is not, of course, the be-all and end-all of the LHC. There is also the question of supersymmetry (SUSY). SUSY is a theory that maintains that every particle in the fermion enclosure of the particle zoo (the quarks, the leptons and the composite particles with an odd number of half-integer spins) has a heavier twin in the boson enclosure (where photons, gluons, and the *W* and *Z* particles currently reside). Thus, the electron (a fermion) has a supersymmetric boson partner, known as a 'selectron', and so on. Theorists love SUSY for her elegance. The LHC will allow us to establish whether SUSY exists or not: even if 'squarks' and 'gluinos' are as heavy as 2.5 TeV, the LHC will find them.

And then there is the question of the extra space dimensions predicted by string theory — that herculean attempt to unify quantum theory and gravitation. For these new dimensions to exist, yet for us to be unaware of them, they must be 'curled up' incredibly small. Theoretically, some might be just big enough to be detected at the LHC through the escape of (gravitational) energy into them.

A speculative laundry list

To me, these three factors — the Higgs particles, supersymmetric particles and new dimensions — are the discoveries most likely to emerge from the first five or so years of LHC operations. But there is a long, more speculative laundry list of objects that might be illuminated by the powerful beams of the LHC. Most of these are speculative in the extreme.

Dark matter origins

Dark matter is one cosmological discovery that has shaken up particle physics, giving rise to many a joint conference with an 'inner space/outer space' theme. The rotational speed of galaxies requires more gravitational 'stuff' than is accounted for by the shining stars. Measurements during the past decade have yielded the information that about 25% of the Universe's mass must be this dark stuff. Neutrinos, which were the initial prime suspects because huge quantities of them were known to have been left over from the Big Bang, are not massive enough. Over time, other exotic candidates — dead suns, black holes and large planets (known as Jupiters) — have been ruled out.

Theorists have supplied us with a plethora of possible solutions, mostly out of their bag of supersymmetric particles. What is known about dark matter is that, first, there is lots of it; second, it does not shine; and, third, it has gravitational force. It is certainly possible that particles will emerge from the collisions of the LHC that will both gladden the hearts of SUSY theorists and account for dark matter.

Dark energy origins

This is, potentially, the elephant in the control room of the LHC. We haven't a clue as to what it is, but we know what it does: it maintains a continuous outward push on the matter of the Universe, sustaining and increasing the expansion rate, and thereby counteracting the gravitational attraction that should be slowing the expansion. It might not be dark, and it might not be energy. But it accounts for more than 70% of the mass of the Universe, so its identification is an important objective. Illumination by the LHC would be a seminal discovery.

Compositeness

Increasingly precise experiments, in the spirit of Ernest Rutherford's scattering experiments on the substructure of gold atoms, have attempted to detect some sort of substructure to the quintessential electron, using progressively more powerful microscopes, each capable of 'seeing' objects smaller than its predecessor: 10^{-18}, 10^{-19}, 10^{-20} cm. This is the maximum scale for an electron's radius (and therefore any internal structure it might have). By necessity, we are now comfortable with the hypothesis that all standard-model particles have zero radius and so no substructure. But this doesn't preclude a future machine detecting a finite size: the higher its energy, the smaller the domain searched.

It is also possible to imagine a sort of substructure that would escape detection by scattering experiments. If one were ever to detect a quark so structured that it could have a higher energy quantum state, and so might absorb energy from the scattered protons, that would be a just-fancy-that moment!

Technicolour

'Technicolour' refers not to the glowing shades in which the fantasy land appears in *The Wizard of Oz* but rather to the quantum field theory postulated as an alternative to the Higgs hypothesis for explaining the masses of the *W* and *Z* particles. Theories that have not yet been confirmed experimentally are judged by their mathematical 'elegance' and their economy in predicting new particles. Supersymmetry predicts a doubling in the total number of particles and must therefore be considered uneconomical. From the point of view of the ambitious experimental physicist desperate to make discoveries, however, SUSY is a godsend. Technicolour predicts a new strong force and a large number of new particles (although fewer than SUSY), whose 'signatures' could stand out above backgrounds of the complex collisions that the LHC will produce.

Strong scattering

One of the wonders of the Higgs hypothesis for high-energy physics is that it cures a particular 'pathology' present in certain predictions: for example, there are infinities in the cross-sections (a measure of the probability of a process) for the scattering of two *W* bosons. The presence of the Higgs would cure this 'disease'. Thus, experiments such as *W*–*W* scattering (I don't know how to do these, but I am sure there are experiments that would include infinite contributions in a Higgs-less theory) should be carried out at the highest energies. If Higgs cannot be discovered, such experiments will be crucial to establish what the missing ingredient that we need to make our theories sensible looks like.

New gauge bosons

Our old force carriers are photons, the *W* and *Z* particles, and gluons. It is strongly assumed that gravitons are 'almost' discovered — although not by suspicious conservatives. Perhaps, a decade into the LHC's operation, our skills in precise analysis of collisions at 14 TeV will have been honed such that we can discover a new force, and so a new boson, predicted by a theorist now in a good high school. At present, our theories don't need such bosons, but that doesn't mean they don't exist.

Right-handed neutrinos

The neutrinos we know and love have less than one millionth the mass of the electron and are left-handed — that is, their spin direction is opposite to their momentum. To be accepted gracefully into current theory, a right-handed neutrino — the spin and momentum of which would be parallel — must be very massive. In addition, we must lose the distinction between the massive neutrino and its antimatter twin. A discovery of such particles at the LHC would be a fantastic step forwards in our quest for a theory of everything. It would, for example, have a bearing on the 'origin of matter' dilemma — that is, why is there a small excess of matter over antimatter, and, by extension, how did we come to be here?

Mini black holes

Black-hole physics deals with the astronomical phenomenon of a massive sun using up its nuclear fuel and eventually collapsing, if it is heavy enough, into a black hole. Of more interest to particle physicists are smaller black holes, left over from the Big Bang, which may well exist in and around our Galaxy. At first pass, even such mini black holes must be much more massive than any imaginable accelerator could reach. But the existence of extra dimensions of finite size, as proposed by string theory, would lower the energy required to produce these hypothetical particles. The idea of the LHC as a mini-black-hole factory is not as worrying as it sounds; they will quickly evaporate through the radiation of energy (Hawking radiation).

What did we leave out?

Theoretical physicists are an imaginative group, and each of these exotic suggestions has its proponents and its naysayers. But the history of the sort of step that the LHC will be making teaches us that, more often than not, a discovery will be made that was not anticipated by theorists. That discovery will change our theories beyond imagination. Fifty years spent investigating the standard model have taught me that, by year ten of the LHC's physics, many an expected and unexpected discovery could well have been celebrated with champagne drunk from styrofoam cups. ■

Leon Lederman is at the Department of Biological, Chemical, and Physical Sciences, Illinois Institute of Technology, 3300 South Federal Street, Chicago, Illinois 60616-3793, USA.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing interests. Correspondence should be addressed to the author (lederman@fnal.gov).

ARTICLES

Generation of germline-competent induced pluripotent stem cells

Keisuke Okita[1], Tomoko Ichisaka[1,2] & Shinya Yamanaka[1,2]

We have previously shown that pluripotent stem cells can be induced from mouse fibroblasts by retroviral introduction of Oct3/4 (also called Pou5f1), Sox2, c-Myc and Klf4, and subsequent selection for *Fbx15* (also called *Fbxo15*) expression. These induced pluripotent stem (iPS) cells (hereafter called Fbx15 iPS cells) are similar to embryonic stem (ES) cells in morphology, proliferation and teratoma formation; however, they are different with regards to gene expression and DNA methylation patterns, and fail to produce adult chimaeras. Here we show that selection for *Nanog* expression results in germline-competent iPS cells with increased ES-cell-like gene expression and DNA methylation patterns compared with Fbx15 iPS cells. The four transgenes (*Oct3/4*, *Sox2*, *c-myc* and *Klf4*) were strongly silenced in Nanog iPS cells. We obtained adult chimaeras from seven Nanog iPS cell clones, with one clone being transmitted through the germ line to the next generation. Approximately 20% of the offspring developed tumours attributable to reactivation of the *c-myc* transgene. Thus, iPS cells competent for germline chimaeras can be obtained from fibroblasts, but retroviral introduction of c-Myc should be avoided for clinical application.

Although ES cells are promising donor sources in cell transplantation therapies[1], they face immune rejection after transplantation and there are ethical issues regarding the usage of human embryos. These concerns may be overcome if pluripotent stem cells can be directly derived from patients' somatic cells[2]. We have previously shown that iPS cells can be generated from mouse fibroblasts by retrovirus-mediated introduction of four transcription factors (Oct3/4 (refs 3, 4), Sox2 (ref. 5), c-Myc (ref. 6) and Klf4 (ref. 7)) and by selection for *Fbx15* expression[8]. Fbx15 iPS cells, however, have different gene expression and DNA methylation patterns compared with ES cells and do not contribute to adult chimaeras. We proposed that the incomplete reprogramming might be due to the selection for *Fbx15* expression, and that by using better selection markers, we might be able to generate more ES-cell-like iPS cells. We decided to use *Nanog* as a candidate of such markers.

Although both *Fbx15* and *Nanog* are targets of Oct3/4 and Sox2 (refs 9–11), Nanog is more tightly associated with pluripotency. In contrast to *Fbx15*-null mice and ES cells that barely show abnormal phenotypes[9], disruption of *Nanog* in mice results in loss of the pluripotent epiblast[12]. *Nanog*-null ES cells can be established, but they tend to differentiate spontaneously[12]. Forced expression of *Nanog* renders ES cells independent of leukaemia inhibitory factor (LIF) for self-renewal[12,13] and confers increased reprogramming efficiency after fusion with somatic cells[14]. These results prompted us to propose that if we use *Nanog* as a selection marker, we might be able to obtain iPS cells displaying a greater similarity to ES cells.

Generation of Nanog iPS cells

To establish a selection system for *Nanog* expression, we began by isolating a bacterial artificial chromosome (BAC, ~200 kilobases) containing the mouse *Nanog* gene in its centre. By using recombineering technology[15,16], we inserted a green fluorescent protein (GFP)-internal ribosome entry site (IRES)-puromycin resistance gene (Puro^r) cassette into the 5′ untranslated region (UTR; Fig. 1a). ES cells that had stably incorporated the modified BAC were positive for GFP, but became negative when differentiation was induced (not shown). By introducing these ES cells into blastocysts, we obtained chimaeric mice and then transgenic mice containing the Nanog-GFP-IRES-Puro^r reporter construct. In transgenic mouse blastocysts, GFP was specifically observed in the inner cell mass (Fig. 1b). In 9.5 days post coitum (d.p.c.) embryos, only migrating primordial germ cells (PGCs) showed GFP signal. In 13.5 d.p.c. embryos, GFP was specifically detected in the genital ridges of both sexes. After removing the brain, visceral tissues and genital ridges, we isolated mouse embryonic fibroblasts (MEFs) from 13.5 d.p.c. male embryos. Flow cytometry analyses showed that these MEFs did not contain GFP-positive cells, whereas ~1% of cells isolated from genital ridges showed GFP signals (Fig. 1c).

Next, we introduced the four previously described factors (Oct3/4, Sox2, Klf4 and the c-Myc mutant c-Myc(T58A)) into Nanog-GFP-IRES-Puro^r MEFs cultured on SNL feeder cells with the use of retroviral vectors. Three, five, or seven days after retroviral infection, we started puromycin selection in ES cell medium. GFP-positive cells first became apparent ~7 days after infection. Twelve days after infection, a few hundred colonies appeared, regardless of the timing of puromycin selection (Fig. 2a). By contrast, no colonies emerged from MEFs transfected with mock DNA. Among puromycin-resistant colonies, ~5% were positive for GFP (Fig. 2b). When the puromycin selection was started at 7 days after infection, we obtained the most GFP-positive colonies. Because we used the GFP-IRES-Puro^r cassette, it is unclear why we obtained GFP-negative colonies. With increased concentrations of puromycin, we obtained fewer GFP-negative colonies (Fig. 2c). With any combination of three of the four factors, we did not obtain any GFP-positive colonies (Supplementary Fig. 1).

By continuing cultivation of these GFP-positive colonies, we obtained cells that were morphologically indistinguishable from ES cells (Fig. 2d). These cells also demonstrated ES-like proliferation, with slightly longer doubling times than that of ES cells (Fig. 3a). Subcutaneous transplantation of these cells into nude mice resulted in tumours that consisted of various tissues of all three germ layers,

[1]Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan. [2]CREST, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan.

indicating that these cells are pluripotent (Fig. 3b and Supplementary Fig. 2). We therefore refer to these cells as Nanog iPS cells in the remainder of this manuscript. Induced pluripotent stem cells were established from Fbx15 β-geo MEFs in parallel and are referred to as Fbx15 iPS cells.

Similarity between Nanog iPS cells and ES cells

Polymerase chain reaction with reverse transcription (RT–PCR) showed that Nanog iPS cells expressed most ES cell marker genes, including *Nanog*, at higher and more consistent levels compared with Fbx15 iPS cells (Fig. 4a). DNA microarray analyses confirmed that Nanog iPS cells had greater ES-cell-like gene expression compared with Fbx15 iPS cells (Fig. 4b). The expression level of *Rex1* (also called *Zfp42*) in Nanog iPS cells was higher compared with Fbx15 iPS cells, but still lower than in ES cells. Thus, Nanog iPS cells show greater gene expression similarity to ES cells (without being identical) than do Fbx15 iPS cells.

RT–PCR showed that Nanog iPS cells have significantly lower expression levels of the four transgenes than Fbx15 iPS cells (Fig. 4c). Real-time PCR confirmed that transgene expression was very low in Nanog iPS cells (Supplementary Fig. 4a–d). In contrast, Southern blot analyses showed similar copy numbers of retroviral integration in Nanog iPS cells and Fbx15 iPS cells (Supplementary Fig. 5). These data indicate that retroviral transgene expression is largely silenced in Nanog iPS cells, as has been shown in ES cells[17]. The expression levels of the transgenes are reversely correlated with *Dnmt3a2* expression, suggesting that *de novo* methyltransferase[18] may be involved in the retroviral silencing observed in iPS cells (Supplementary Fig. 6).

Bisulphite genomic sequencing analyses also revealed similarities between Nanog iPS cells and ES cells (Fig. 5). The promoter regions of *Nanog*, *Oct3/4* and *Fbx15* were largely unmethylated in Nanog iPS cells. This is in marked contrast to Fbx15 iPS cells in which the promoters of *Nanog* and *Oct3/4* were only partially unmethylated[8]. Differentially methylated regions of imprinting genes *H19* and *Igf2r* were partially methylated in Nanog iPS cells. During PGC development, imprinting is erased by 12.5 d.p.c.[19–21]. The loss of imprinting is maintained in embryonic germ cells derived from 12.5 d.p.c. PGCs[22] and cloned embryos derived from 12.5–16.5 d.p.c. PGCs[23,24]. ES cells, by contrast, showed normal imprinting patterns[25]. Thus, Nanog iPS cells show greater similarity in the methylation patterns of imprinting genes to ES cells than to embryonic germ cells.

Simple sequence length polymorphism (SSLP) analyses showed that Nanog iPS cells are largely of the DBA background but also have some contribution from the C57BL/6 and 129S4 backgrounds (Supplementary Fig. 3). This result is consistent with the genetic background of the MEFs, which was 75% DBA, 12.5% C57BL/6

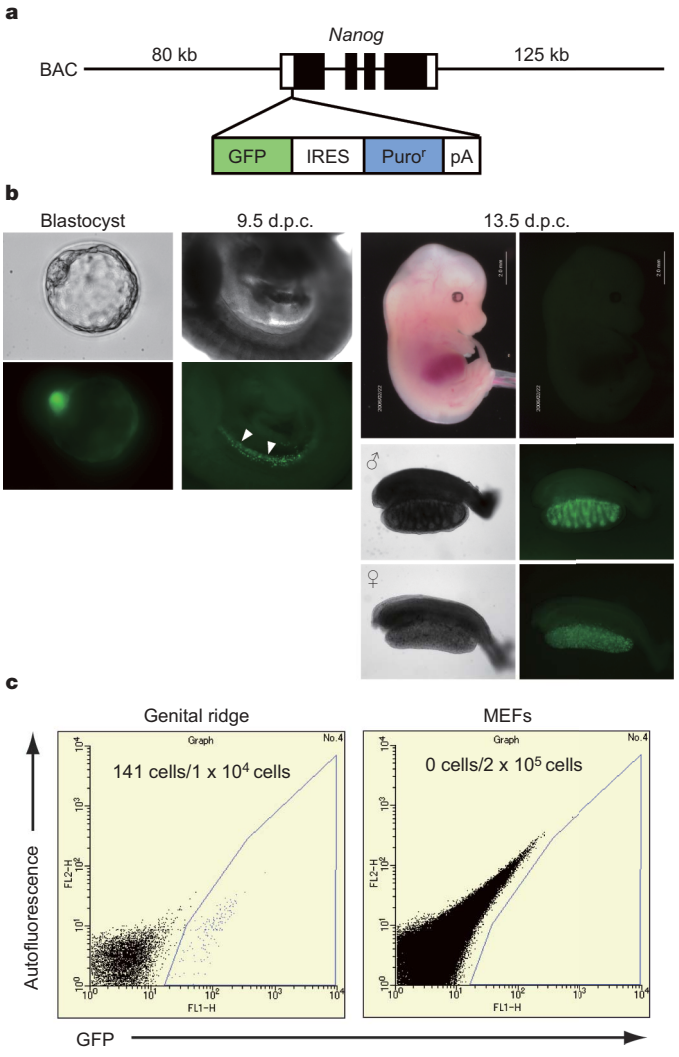


Figure 1 | Nanog-GFP-IRES-Puro^r transgenic mice. **a**, Modified BAC construct. White boxes indicate the 5′ and 3′ UTRs of the mouse *Nanog* gene. Black boxes indicate the open reading frame. **b**, GFP expression in Nanog-GFP transgenic mouse embryos. Whole embryos (top panels) and isolated genital ridges (bottom panels) from 13.5 d.p.c. mice are shown. **c**, Histogram showing GFP fluorescence in cells isolated from genital ridges of a 13.5 d.p.c. Nanog-GFP transgenic mouse embryo (left) or in MEFs isolated from the same embryo (right).

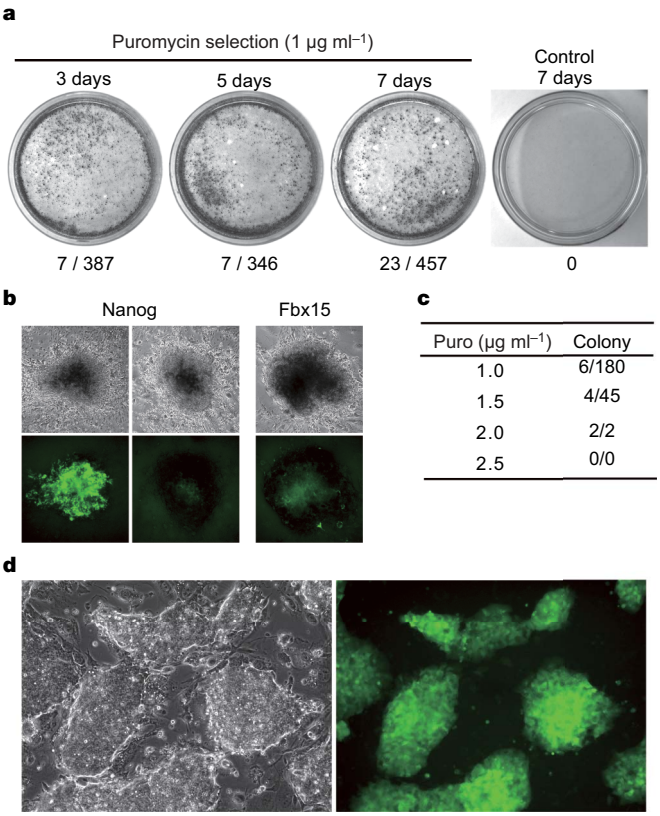


Figure 2 | Generation of iPS cells from MEFs of Nanog-GFP-IRES-Puro^r transgenic mice. **a**, Puromycin-resistant colonies. Puromycin selection was initiated at 3, 5, or 7 days after retroviral transduction. Numbers indicate GFP-positive colonies/total colonies. **b**, GFP fluorescence in resulting colonies. Phase contrast (top row) and fluorescence (bottom row) micrographs are shown. iPS cells were also generated from Fbx15-βgeo knockin MEFs. **c**, Effect of increasing concentrations of puromycin. Numbers of GFP-positive colonies/total colonies are shown on the right. **d**, Morphology of established Nanog iPS cells (clone 20D17). Phase contrast (left) and fluorescence (right) micrographs are shown.

and 12.5% 129S4. This result also confirms that Nanog iPS cells are not a contamination of ES cells that exists in our laboratory, which are either pure 129S4 or C57BL/6.

We next compared the stability of Nanog iPS cells and Fbx15 iPS cells (Supplementary Fig. 7). Cells were cultivated in the presence of the selection drug for up to 22–26 passages. Morphologically, we did not observe significant changes over the long-term culture course. However, RT–PCR showed that Fbx15 iPS cells lost the expression of ES cell marker genes after prolonged culture. By contrast, Nanog iPS cells maintained relatively high expression levels of the ES cell marker genes. These data demonstrate that Nanog iPS cells are more stable than Fbx15 iPS cells.

We also compared the induction efficiency of Nanog iPS cells and Fbx15 iPS cells. In independent experiments, we obtained 4–125 GFP-positive colonies from 8×10^5 Nanog-reporter MEFs transfected with the four transcription factors. Because ~50% of transfected MEFs are supposed to express all four factors[8], the induction efficiency is approximately 0.001–0.03%. In contrast, from the same number of Fbx15-reporter MEFs, we obtained 47–1,800 G418-resistant colonies, with the induction efficiency of approximately 0.01–0.5%. Thus, the efficiency of Nanog iPS cell induction is approximately one-tenth that of Fbx15 iPS cells.

We then compared the responses of Nanog iPS cells and Fbx15 iPS cells to LIF or retinoic acid (Supplementary Fig. 8). As we have shown previously[8], Fbx15 iPS cells do not remain undifferentiated when cultured without feeder cells, even in the presence of LIF. Furthermore, Fbx15 iPS cells formed compact colonies when cultured without feeder cells in the presence of retinoic acid. In contrast, LIF maintained the undifferentiated state of Nanog iPS cells cultured without feeder cells. Retinoic acid induced the differentiation of Nanog iPS cells. Thus, Nanog iPS cells are similar to ES cells in their response to LIF and retinoic acid.

Initially we used the T58A mutant of c-Myc to induce Nanog iPS cells. We also tested wild-type c-Myc for Nanog iPS cell induction. We obtained a similar number of colonies with both wild-type c-Myc and the T58A mutant. Nanog iPS cells established with wild-type c-Myc were indistinguishable from those established with the T58A mutant with regards to morphology, gene expression (analysed via microarrays), teratoma formation (Supplementary Fig. 2) and stability under puromycin selection (Supplementary Fig. 9). Without puromycin selection, Nanog iPS cells induced by wild-type c-Myc were more stable (Supplementary Fig. 9).

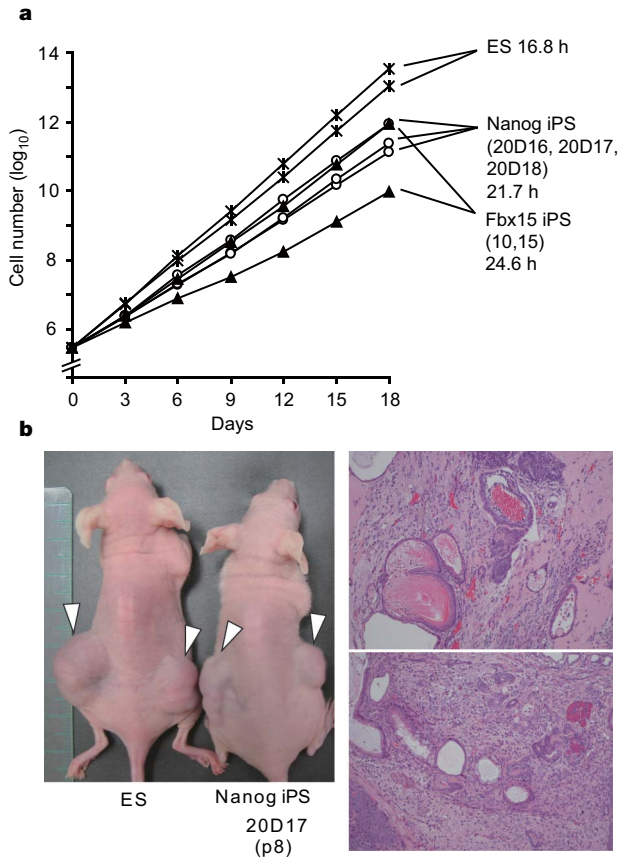


Figure 3 | Characterization of Nanog iPS cells. **a**, Proliferation. ES cells, Nanog iPS cells (clones 20D16, 20D17 and 20D18) and Fbx15 iPS cells (clones 10 and 15) were passaged every 3 days (3×10^5 cells per each well of a 6-well plate). Calculated doubling times are indicated. **b**, Teratomas. ES cells or Nanog iPS cells (clone 20D17, 1×10^6 cells) were subcutaneously transplanted into nude mice. After 8 weeks, teratomas were photographed (left) and analysed histologically with haematoxylin and eosin staining (right).

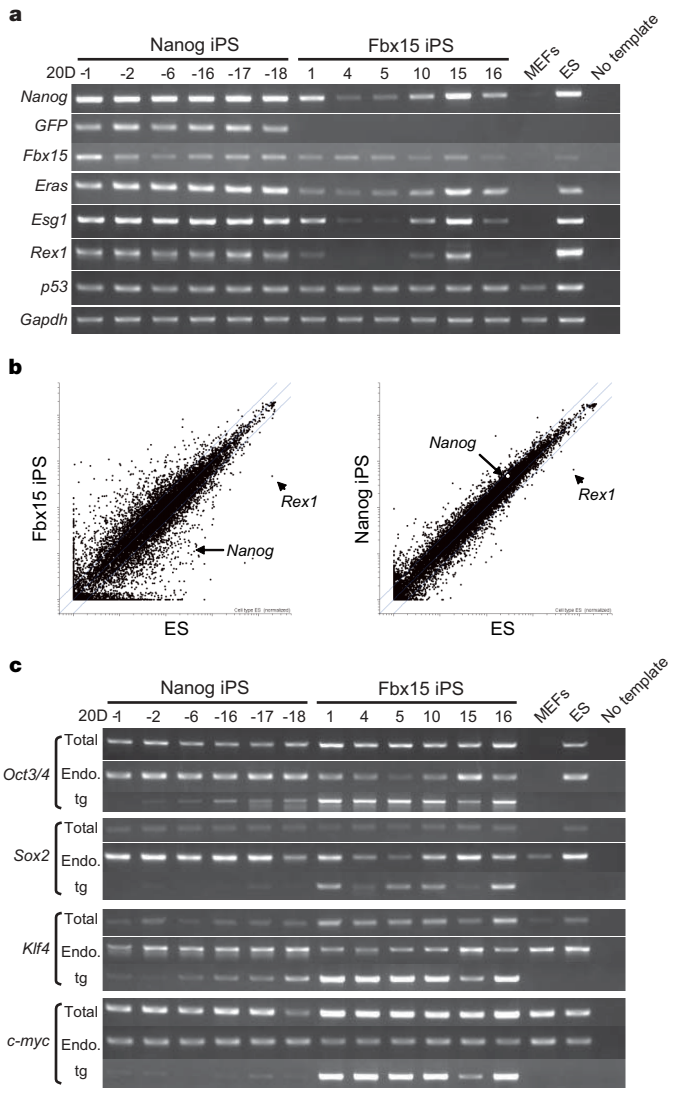


Figure 4 | Gene expression in Nanog iPS cells. **a**, RT–PCR. Total RNA was isolated from six clones of Nanog iPS cells (clones 20D1, 20D2, 20D6, 20D16, 20D17 and 20D18), six clones of Fbx15 iPS cells (clones 1, 4, 5, 10, 15 and 16), MEFs and ES cells. **b**, Scatter plots showing comparison of global gene expression between ES cells and Nanog iPS cells (right), and between ES cells and Fbx15 iPS cells (left), as determined by DNA microarrays. **c**, Expression levels of the four transcription factors. Total RNA was isolated from six clones of Nanog iPS cells (clones 20D1, 20D2, 20D6, 20D16, 20D17 and 20D18), six clones of Fbx15 iPS cells (clones 1, 4, 5, 10, 15 and 16), MEFs and ES cells. RT–PCR analyses were performed with primers that amplified the coding regions of the four factors (Total), endogenous transcripts only (Endo.), and transgene transcripts only (tg).

Germline chimaeras from Nanog iPS cells

We next examined the ability of Nanog iPS cells to produce adult chimaeras. We injected 15–20 male Nanog iPS cells (five clones with the T58A mutant and three with wild-type c-Myc) into C57BL/6-derived blastocysts, which we then transplanted into the uteri of pseudo-pregnant mice. We obtained adult chimaeras from seven clones (four clones with the T58A mutant and three with wild-type c-Myc) as determined by coat colour (Fig. 6a and Supplementary Table 1). SSLP analyses showed that Nanog iPS cells contributed to various organs, with the level of chimaerism ranging from 10% to 90%. Chimeras from clone 20D17 showed highest iPS cell contribution in the testes. From clone 20D18, we obtained only a few non-chimaeric pups from infected blastocysts; thus, whether this clone has competency for producing adult chimaeras remained to be determined. These data demonstrate that most Nanog iPS clones are competent for adult chimaeric mice.

We then crossed three of the chimaeras from clone 20D17—for which the highest iPS cell contribution was in the testes—with C57BL/6 females. Whereas all F_1 mice showed black coat colour, all contained retroviral integration of the four transcription factors and approximately half contained the GFP-IRES-Puro^r cassette (Fig. 6b), indicating germline transmission. Furthermore, approximately half of the F_2 mice born from F_1 intercrosses showed agouti coat colour, confirming germline transmission of Nanog-iPS-20D17 (Fig. 6c).

We also examined germline competency for two other clones that produced adult chimaeras. In one chimaeric mouse from Nanog-iPS-38C2 cell line, PCR analysis detected iPS cell contribution in isolated spermatozoa (Fig. 6d), suggesting that germline competency is not confined to clone 20D17. However, the iPS cell contribution to sperm of clone 38C2 is much smaller than that of clone 20D17, and no iPS-cell-derived offspring were found for 119 mice born from the cross between the 38C2 chimaera and C57BL/6 female mice. Most male mice with a high degree of chimaerism from the Nanog-iPS-38D2 cell line showed small testes and aspermatogenesis (Supplementary Fig. 13). The testes of some chimaeras from Nanog-iPS-38D2 contained mature sperm, but no iPS cell contribution was detected by PCR (Fig. 6d).

Tumour formation by *c-myc* reactivation

Out of 121 F_1 mice (aged 8–41 weeks) derived from the Nanog-iPS-20D17 cell line, 24 died or were killed because of weakness, wheezing or paralysis. Necropsy of 17 mice identified neck tumours (Supplementary Fig. 10) in 13 mice and other tumours in five mice, including two mice with neck tumours. Histological examination of one neck tumour showed that it was a ganglioneuroblastoma with follicular carcinoma of the thyroid gland (not shown). In these tumours, retroviral expression of *c-myc*, but not *Oct3/4*, *Sox2*, or *Klf4*, is reactivated (Supplementary Fig. 11). In contrast, transgene expression of all four transcription factors remained low in normal tissues, except for *c-myc* in muscle in one mouse (Supplementary Fig. 12). These data indicate that reactivation of *c-myc* retrovirus is attributable to tumour formation.

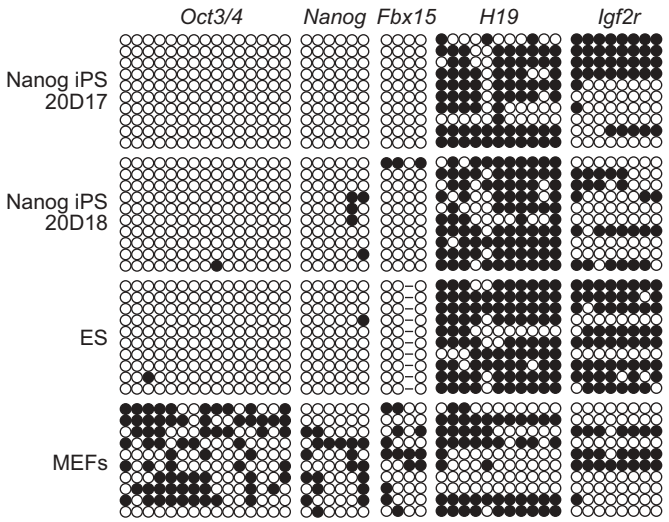


Figure 5 | DNA methylation of ES-cell-specific genes and imprinting genes. White circles indicate unmethylated CpG dinucleotides, whereas black circles indicate methylated CpG dinucleotides.

Discussion

Our results demonstrate that Nanog selection allows the generation of high-quality iPS cells that are comparable to ES cells in morphology, proliferation, teratoma formation, gene expression and competency for adult chimaeras. Nearly all Nanog iPS clones showed these properties, indicating that Nanog is a major determinant of quality in cellular pluripotency. However, germline competence was variable among Nanog iPS clones, indicating the existence of other important determinants of germline competency in addition to Nanog. The high quality of Nanog iPS cells underscores the possibility of using this technology to generate patient-specific pluripotent stem cells. In a separate study, we found that germline-competent iPS cells can also be obtained from adult mouse somatic cells (T. Aoi and S.Y., unpublished data). The current study, however, also reveals that reactivation of *c-myc* retrovirus may result in tumour formation. There may be ways to overcome this problem. Strong silencing of the four retroviruses in Nanog iPS cells indicates that they are only required for the induction, but not the maintenance, of pluripotency. Therefore, the retrovirus-mediated system might be

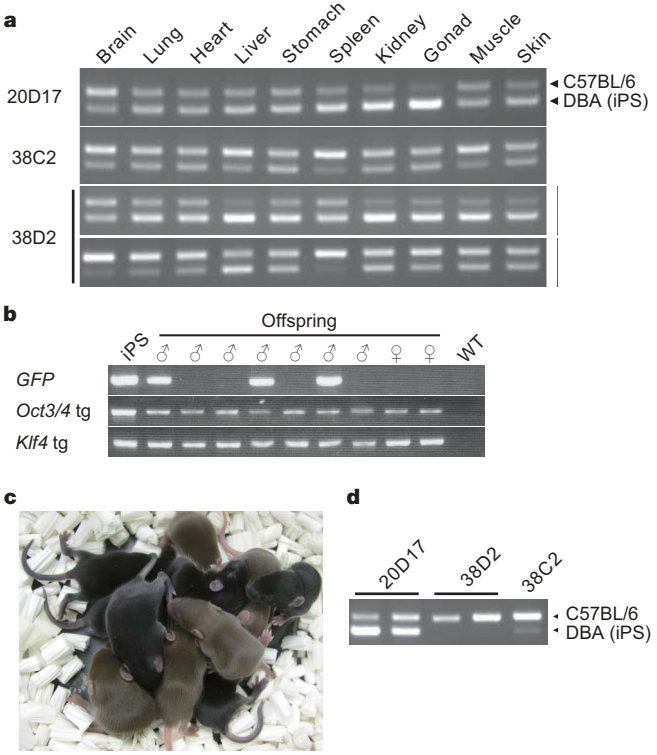


Figure 6 | Germline chimaeras from Nanog iPS cells. a, Tissue distribution of iPS cells in chimaeras. Genomic DNA was isolated from the indicated organs of chimaeras derived from three Nanog iPS cell clones (20D17, 38C2 and 38D2). SSLP analyses were performed for D6Mit15. **b**, PCR analyses showing the presence of the GFP cassette and retroviral transgenes in F_1 mice obtained from the intercross between a chimaeric male and a C57BL/6 female. **c**, Coat colours of F_2 mice obtained from F_1 intercrosses. **d**, Sperm contribution of iPS cells in chimaeric mice. Spermatozoa were isolated from the epididymides of chimaeric mice derived from three Nanog iPS cell clones (20D17, 38D2 and 38C2). iPS cell contribution was determined by SSLP of D6Mit15.

eventually replaced by transient expression, such as the adenovirus-mediated system. Alternatively, high-throughput screening of chemical libraries might identify small molecules that can replace the four genes. These are crucial research areas in order to apply iPS cells to regenerative medicine.

We found that the efficiency of Nanog iPS cell induction is less than 0.1%. The low efficiency suggests that the origin of iPS cells might be rare stem cells co-existing in MEF culture. Alternatively, activation of additional genes by retroviral integration might be required for iPS cell generation in addition to the four transcription factors. This is relevant to the fact that we have been able to obtain iPS cells only with retroviral transduction. Identification of such factor(s) may lead to the generation of iPS cells with higher efficiency, and without the need for retroviruses.

METHODS SUMMARY

To generate Nanog-reporter mice, we isolated a BAC clone containing the mouse *Nanog* gene in its centre. By using the RED/ET recombination technique (Gene Bridges), we inserted a GFP-IRES-Puro^r cassette into the 5′ UTR of the mouse *Nanog* gene. We introduced the modified BAC into RF8 ES cells by electroporation[26]. We then microinjected transgenic ES cells into C57BL/6 blastocysts to generate Nanog-reporter mice containing the modified BAC. MEFs were isolated from 13.5 d.p.c. male embryos after removing genital ridges. Generation of Nanog iPS cells was performed as described[8], except that puromycin was used instead of G418 as a selection antibiotic. Retroviruses (pMXs) were generated with Plat-E packaging cells[27]. RF8 ES cells[26] and iPS cells were cultured on SNL feeder cells[28]. Analyses of iPS cells, such as RT–PCR, real-time PCR, bisulphite genomic sequencing, SSLP analyses, DNA microarrays, teratoma formation, and microinjection into C57BL/6 blastocysts, were performed as described[8]. Contribution of iPS cells in chimaeric mice was determined by PCR for the SSLP marker D6Mit15.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 February; accepted 22 May 2007.
Published online 6 June 2007.

1. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282,** 1145–1147 (1998).
2. Hochedlinger, K. & Jaenisch, R. Nuclear reprogramming and pluripotency. *Nature* **441,** 1061–1067 (2006).
3. Niwa, H., Miyazaki, J. & Smith, A. G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genet.* **24,** 372–376 (2000).
4. Nichols, J. *et al.* Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95,** 379–391 (1998).
5. Avilion, A. A. *et al.* Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* **17,** 126–140 (2003).
6. Cartwright, P. *et al.* LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* **132,** 885–896 (2005).
7. Li, Y. *et al.* Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. *Blood* **105,** 635–637 (2005).
8. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
9. Tokuzawa, Y. *et al.* Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. *Mol. Cell. Biol.* **23,** 2699–2708 (2003).
10. Kuroda, T. *et al.* Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. *Mol. Cell. Biol.* **25,** 2475–2485 (2005).
11. Rodda, D. J. *et al.* Transcriptional regulation of Nanog by OCT4 and SOX2. *J. Biol. Chem.* **280,** 24731–24737 (2005).
12. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113,** 631–642 (2003).
13. Chambers, I. *et al.* Functional expression cloning of nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113,** 643–655 (2003).
14. Silva, J., Chambers, I., Pollard, S. & Smith, A. Nanog promotes transfer of pluripotency after cell fusion. *Nature* **441,** 997–1001 (2006).
15. Copeland, N. G., Jenkins, N. A. & Court, D. L. Recombineering: a powerful new tool for mouse functional genomics. *Nature Rev. Genet.* **2,** 769–779 (2001).
16. Testa, G. *et al.* Engineering the mouse genome with bacterial artificial chromosomes to create multipurpose alleles. *Nature Biotechnol.* **21,** 443–447 (2003).
17. Cherry, S. R., Biniszkiewicz, D., van Parijs, L., Baltimore, D. & Jaenisch, R. Retroviral expression in embryonic stem cells and hematopoietic stem cells. *Mol. Cell. Biol.* **20,** 7419–7426 (2000).
18. Chen, T., Ueda, Y., Xie, S. & Li, E. A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. *J. Biol. Chem.* **277,** 38746–38754 (2002).
19. Davis, T. L., Yang, G. J., McCarrey, J. R. & Bartolomei, M. S. The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. *Hum. Mol. Genet.* **9,** 2885–2894 (2000).
20. Sato, S., Yoshimizu, T., Sato, E. & Matsui, Y. Erasure of methylation imprinting of Igf2r during mouse primordial germ-cell development. *Mol. Reprod. Dev.* **65,** 41–50 (2003).
21. Brandeis, M. *et al.* The ontogeny of allele-specific methylation associated with imprinted genes in the mouse. *EMBO J.* **12,** 3669–3677 (1993).
22. Labosky, P. A., Barlow, D. P. & Hogan, B. L. Mouse embryonic germ (EG) cell lines: transmission through the germline and differences in the methylation imprint of insulin-like growth factor 2 receptor (Igf2r) gene compared with embryonic stem (ES) cell lines. *Development* **120,** 3197–3204 (1994).
23. Kato, Y. *et al.* Developmental potential of mouse primordial germ cells. *Development* **126,** 1823–1832 (1999).
24. Lee, J. *et al.* Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. *Development* **129,** 1807–1817 (2002).
25. Geijsen, N. *et al.* Derivation of embryonic germ cells and male gametes from embryonic stem cells. *Nature* **427,** 148–154 (2004).
26. Meiner, V. L. *et al.* Disruption of the acyl-CoA:cholesterol acyltransferase gene in mice: evidence suggesting multiple cholesterol esterification enzymes in mammals. *Proc. Natl Acad. Sci. USA* **93,** 14041–14046 (1996).
27. Morita, S., Kojima, T. & Kitamura, T. Plat-E: an efficient and stable system for transient packaging of retroviruses. *Gene Ther.* **7,** 1063–1066 (2000).
28. McMahon, A. P. & Bradley, A. The Wnt-1 (int-1) proto-oncogene is required for development of a large region of the mouse brain. *Cell* **62,** 1073–1085 (1990).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Takahashi, M. Nakagawa and T. Aoi for scientific discussion; M. Maeda for histological analyses; M. Narita, J. Iida, H. Miyachi and S. Kitano for technical assistance; and R. Kato, R. Iyama and Y. Ohuchi for administrative assistance. We also thank T. Kitamura for Plat-E cells and pMXs retroviral vectors, and R. Farese for RF8 ES cells. This study was supported in part by a grant from the Uehara Memorial Foundation, the Program for Promotion of Fundamental Studies in Health Sciences of NIBIO, a grant from the Leading Project of MEXT, and Grants-in-Aid for Scientific Research of JSPS and MEXT (to S.Y.). K.O. is a JSPS research fellow.

Author Contributions K.O. conducted most of the experiments in this study. T.I. performed manipulation of mouse embryos to generate Nanog-GFP transgenic mice. T.I. also maintained the mouse lines. S.Y. designed and supervised the study, and prepared the manuscript. S.Y. also performed computer analyses of DNA microarray data.

Author Information The microarray data are deposited in GEO under accession number GSE7841. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.Y. (yamanaka@frontier.kyoto-u.ac.jp).

METHODS

Cell culture. RF8 ES cells (129S4 background)[26] and iPS cells were maintained in ES medium (DMEM containing 15% FCS, 1× NEAA, 1 mM sodium pyruvate, 5.5 mM 2-ME, 50 units ml^{-1} penicillin and 50 μg ml^{-1} streptomycin) on feeder layers of mitomycin-C-treated SNL cells into which we had stably incorporated the puromycin-resistance gene. As a source of LIF, we used conditioned medium from Plat-E cell cultures that had been transduced with a LIF-expressing vector. Plat-E cells[27], which were also used to produce retroviruses, were maintained in DMEM containing 10% FCS, 50 units ml^{-1} penicillin, 50 μg ml^{-1} streptomycin, 1 μg ml^{-1} puromycin and 10 μg ml^{-1} blasticidin S.

For MEF isolation, we used 13.5 d.p.c. male embryos. After the removal of the head, visceral tissues and gonads, the remaining bodies were washed and dissociated with trypsin. Ten-million cells were plated on each gelatin-coated 100-mm dish and incubated at 37 °C with 5% CO_2. The next day, floating cells were removed by washing with PBS. In this study, MEFs were used within passage 5 to avoid replicative senescence.

BAC modification. A mouse BAC clone containing the *Nanog* gene, RP23-117I23, was purchased from BACPAC Resources. Modification of the BAC was performed using the RED/ET recombination technique (Gene Bridges), which is based on homologous recombination using inducible RecET recombination machinery. The reporter cassette was made by ligating a GFP-IRES-$Puro^r$ fragment with a PGK-hygro-FRT cassette (Gene Bridges) that contains a prokaryotic promoter and hygromycin-resistance gene flanked by FRT sites. Homology arms for the *Nanog* 5′ UTR were attached to both ends of the reporter cassette by PCR amplification using the following primers: BAC-NanogGFP-F (5′-TTTGCATTAGACATTTAACTCTTCTTTCTATGATCTTTCCTTCTAGACACGCCACCATGGTGAGCAAGGGCGAG-3′) and BAC-NanogR (5′-GCGAGGGAAGGGATTTCTGAAAAGGTTTTAGGCAACAACCAAAAAACTCACTGGCAGTTTATGGCGGGCGTCCT-3′). *Escherichia coli* carrying the BAC was transformed with a RED/ET expression plasmid, pSC101-BAD-gbaA, and recombination with the reporter cassette was subsequently induced. Successfully recombined colonies were identified by screening for hygromycin resistance followed by PCR analysis to ensure homologous recombination. The hygromycin cassette was excised by transformation into a Flp-recombinase expression bacterium, 294-Flp (Gene Bridges).

Establishment of Nanog-reporter mice. The modified BAC was linearized by *Not*I digestion, and 10 μg of the DNA was introduced into RF8 ES cells by electroporation[26]. After 2 days, selection was started with 1.5 μg ml^{-1} puromycin. Resistant colonies were picked after 9 days of selection. Four genomic integrated clones were used for blastocyst injection, and we established two lines of Nanog reporter mice: 2A2 and 2C1. Both mice exhibited the same expression pattern. Mice from the 2A2 line were used for iPS cell induction.

iPS cell induction. iPS induction was performed as described previously[8] with some modifications. Briefly, MEFs were isolated from 13.5 d.p.c. embryos from Nanog-reporter male mice (50% DBA, 25% C57BL/6 and 25% 129S4) and female wild-type mice (DBA). Plat-E cells were seeded at 8×10^6 cells per 100-mm dish. On the next day, 9 μg of pMXs-based retroviral vectors for *Oct3/4*, *Sox2*, *Klf4*, or *c-Myc* were independently introduced into Plat-E cells using 27 μl of FuGENE 6 transfection reagent. After 24 h, the medium was replaced with 10 ml of DMEM containing 10% FCS. MEFs were seeded at 8×10^5 cells per 100-mm dish covered by feeder cells. On the next day, virus-containing supernatants from these Plat-E cultures were recovered and filtered through a 0.45-μm cellulose acetate filter. Equal volumes of the supernatants were mixed and supplemented with polybrene at the final concentration of 4 μg ml^{-1}. MEFs were incubated in the virus/polybrene-containing supernatants for 24 h. Three days after infection, the medium was changed with ES medium supplemented with LIF. For Fbx15 iPS cell selection, we added G418 (Geneticin from Invitrogen) at a final concentration of 0.3 mg ml^{-1}. For Nanog iPS cells, we added puromycin (Sigma) at a final concentration of 1.5 μg ml^{-1}, unless indicated otherwise. Established iPS cells were maintained in the presence of the corresponding selection drug. Teratoma formation, RT–PCR analysis, Bisulphite sequence and SSLP analysis were performed as previously described[8].

DNA microarray. Total RNA from *Fbx15*-null MEFs (duplicate), *Fbx15*-null ES cells (duplicate), Fbx15 iPS cells (clone MEF4-7[8], duplicate), or Nanog iPS cells (clones 20D -2, 16, 17, and 18) were labelled with Cy3. Samples were hybridized to a Mouse Oligo Microarray (Agilent) according to the manufacturer's protocol. Arrays were scanned with a G2565BA Microarray Scanner System (Agilent). Data were analysed using GeneSprings GX software (Agilent). We excluded genes for which their value fluctuated more than twofold between duplicated analyses.

ARTICLES

In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state

Marius Wernig[1]*, Alexander Meissner[1]*, Ruth Foreman[1,2]*, Tobias Brambrink[1]*, Manching Ku[3]*, Konrad Hochedlinger[1]†, Bradley E. Bernstein[3,4,5] & Rudolf Jaenisch[1,2]

Nuclear transplantation can reprogramme a somatic genome back into an embryonic epigenetic state, and the reprogrammed nucleus can create a cloned animal or produce pluripotent embryonic stem cells. One potential use of the nuclear cloning approach is the derivation of 'customized' embryonic stem (ES) cells for patient-specific cell treatment, but technical and ethical considerations impede the therapeutic application of this technology. Reprogramming of fibroblasts to a pluripotent state can be induced *in vitro* through ectopic expression of the four transcription factors Oct4 (also called Oct3/4 or Pou5f1), Sox2, c-Myc and Klf4. Here we show that DNA methylation, gene expression and chromatin state of such induced reprogrammed stem cells are similar to those of ES cells. Notably, the cells—derived from mouse fibroblasts—can form viable chimaeras, can contribute to the germ line and can generate live late-term embryos when injected into tetraploid blastocysts. Our results show that the biological potency and epigenetic state of *in-vitro*-reprogrammed induced pluripotent stem cells are indistinguishable from those of ES cells.

Epigenetic reprogramming of somatic cells into ES cells has attracted much attention because of the potential for customized transplantation therapy, as cellular derivatives of reprogrammed cells will not be rejected by the donor[1,2]. Thus far, somatic cell nuclear transfer and fusion of fibroblasts with ES cells have been shown to promote the epigenetic reprogramming of the donor genome to an embryonic state[3–5]. However, the therapeutic application of either approach has been hindered by technical complications as well as ethical objections[6]. Recently, a major breakthrough was reported whereby expression of the transcription factors Oct4, Sox2, c-Myc and Klf4 was shown to induce fibroblasts to become pluripotent stem cells (designated as induced pluripotent stem (iPS) cells), although with a low efficiency[7]. The iPS cells were isolated by selection for activation of *Fbx15* (also called *Fbxo15*), which is a downstream gene of *Oct4*. This important study left a number of questions unresolved: (1) although iPS cells were pluripotent they were not identical to ES cells (for example, iPS cells injected into blastocysts generated abnormal chimaeric embryos that did not survive to term); (2) gene expression profiling revealed major differences between iPS cells and ES cells; (3) because the four transcription factors were transduced by constitutively expressed retroviral vectors it was unclear why the cells could be induced to differentiate and whether continuous vector expression was required for the maintenance of the pluripotent state; and (4) the epigenetic state of the endogenous pluripotency genes *Oct4* and *Nanog* was incompletely reprogrammed, raising questions about the stability of the pluripotent state.

Here we used activation of the endogenous *Oct4* or *Nanog* genes as a more stringent selection strategy for the isolation of reprogrammed cells. We infected fibroblasts with retroviral vectors transducing the four factors, and selected for the activation of the endogenous *Oct4* or *Nanog* genes. Positive colonies resembled ES cells and assumed an epigenetic state characteristic of ES cells. When injected into blastocysts the reprogrammed cells generated viable chimaeras and contributed to the germ line. Our results establish that somatic cells can be reprogrammed to a pluripotent state that is similar, if not identical, to that of normal ES cells.

Selection of fibroblasts for *Oct4* or *Nanog* activation

Using homologous recombination in ES cells we generated mouse embryonic fibroblasts (MEFs) and tail-tip fibroblasts (TTFs) that carried a neomycin-resistance marker inserted into either the endogenous *Oct4* (Oct4-neo) or *Nanog* locus (Nanog-neo) (Fig. 1a). These cultures were sensitive to G418, indicating that the *Oct4* and *Nanog* loci were, as expected, silenced in somatic cells. These MEFs or TTFs were infected with Oct4-, Sox2-, c-Myc- and Klf4-expressing retroviral vectors and G418 was added to the cultures 3, 6 or 9 days later. The number of drug-resistant colonies increased substantially when analysed at day 20 (Fig. 1i). Most colonies had a flat morphology (Fig. 1h, right) and between 11% and 25% of the colonies were 'ES-like' (Fig. 1h, left) when selection was applied early (Fig. 1k), a percentage that increased at later time points. At day 20, ES-like colonies were picked, dissociated and propagated in G418-containing media. They gave rise to ES-like cell lines (designated as Oct4 iPS or Nanog iPS cells, respectively) that could be propagated without drug selection, displayed homogenous Nanog, SSEA1 and alkaline phosphatase expression (Fig. 1b–g and Supplementary Figs 1 and 5), and formed undifferentiated colonies when seeded at clonal density on gelatin-coated dishes (see inset in Fig. 1b). Four out of five analysed lines had a normal karyotype (Supplementary Table 1).

Although the timing and appearance of colonies were similar between the Oct4 and Nanog selection, we noticed pronounced quantitative differences between the two selection strategies: whereas Oct4-selected MEF cultures had 3- to 10-fold fewer colonies, the fraction of ES-like colonies was 2- to 3-fold higher than in Nanog-selected cultures. Accordingly, approximately four times more Oct4-selected ES-like colonies gave rise to stable and homogenous iPS cell

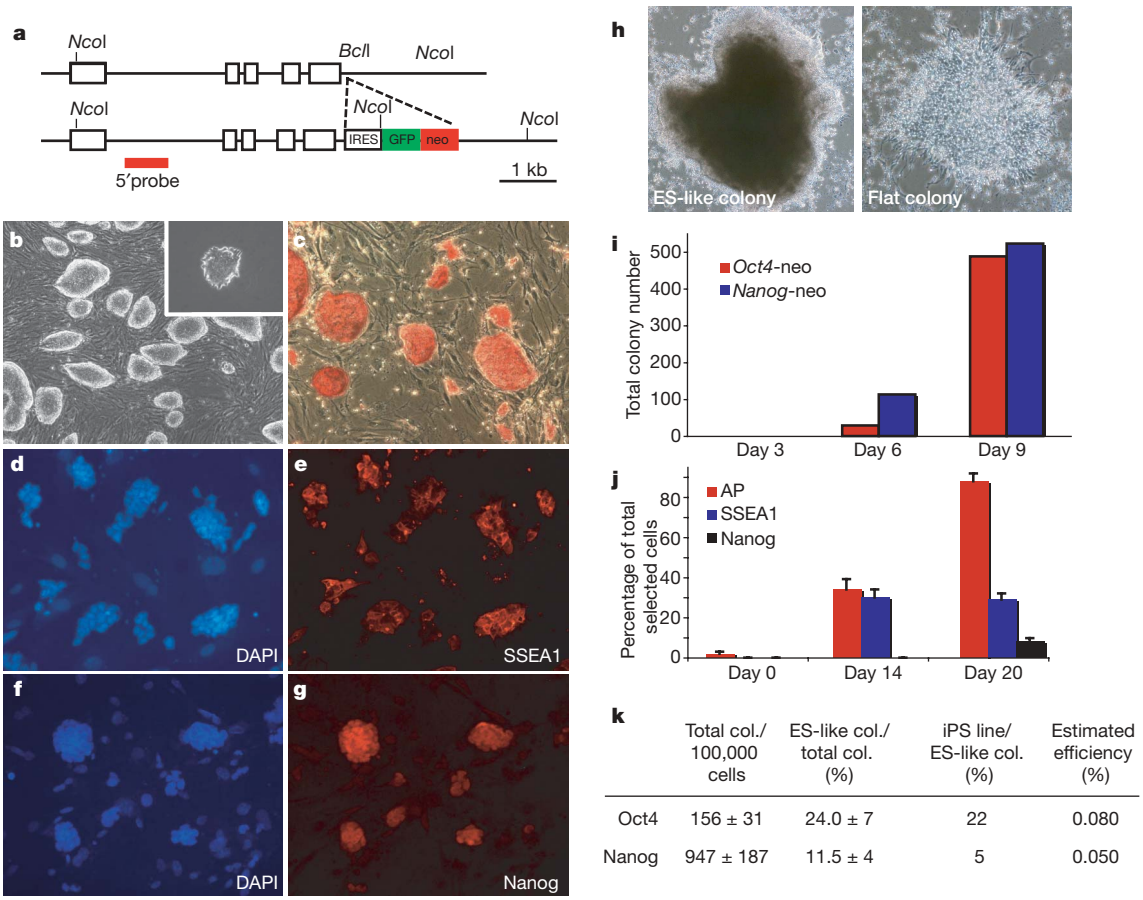
[1]Whitehead Institute for Biomedical Research and [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [3]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [4]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [5]Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. †Present address: Center for Regenerative Medicine and Cancer Center, Massachusetts General Hospital, Harvard Medical School and Harvard Stem Cell Institute, Boston, Massachusetts 02414, USA.
*These authors contributed equally to this work.

lines compared with Nanog-selected ES-like colonies (Fig. 1k). This suggests that although the *Nanog* locus was easier to activate, a higher fraction of the drug-resistant colonies in Oct4-neo cultures was reprogrammed to a pluripotent state. Therefore, the overall estimated efficiency of 0.05–0.1% to establish iPS cell lines from MEFs was similar between Oct4 selection and Nanog selection, despite the larger number of total Nanog-neo resistant colonies (Fig. 1k). Next we investigated the time course of reprogramming by studying the fraction of alkaline-phosphatase-, SSEA1- and Nanog-positive cells in Oct4-selected MEF cultures. Fourteen days after infection some cells had already initiated alkaline phosphatase activity and SSEA1 expression, but lacked detectable amounts of Nanog protein (Fig. 1j), whereas by day 20, alkaline phosphatase and SSEA1 expression had increased and ~8% of the cells were Nanog-positive. Thus, the reprogramming induced by the four transcription factors (Oct4, Sox2, c-Myc and Klf4) is a gradual and slow process.

Expression and DNA methylation

To characterize the reprogrammed cells on a molecular level we used quantitative polymerase chain reaction with reverse transcription (qRT–PCR) to measure the expression of ES-cell- and fibroblast-specific genes. Figure 2a shows that in Oct4 iPS cells the total level of Nanog and Oct4 was similar to that in ES cells but decreased on differentiation to embryoid bodies. MEFs did not express either gene. Using specific primers for endogenous or total *Sox2* transcripts showed that most *Sox2* transcripts originated from the endogenous locus rather than the viral vector (Fig. 2b). In contrast, *Hoxa9* and *Zfpm2* were highly expressed in MEFs but were expressed at very low levels in iPS or ES cells (Fig. 2c). Western blot analysis showed that multiple iPS clones expressed Nanog and Oct4 proteins at similar levels compared to ES cells (Fig. 2d). Finally, we used microarray technology to compare gene expression patterns on a global level. Figure 2f shows that the iPS cells clustered with ES cells in contrast to wild-type or donor MEFs.

To investigate the DNA methylation level of the *Oct4* and *Nanog* promoters we performed bisulphite sequencing and combined bisulphite restriction analysis (COBRA) with DNA isolated from ES cells, iPS cells and MEFs. As shown in Fig. 2g, both loci were demethylated in ES and iPS cells and fully methylated in MEFs. To assess whether the maintenance of genomic imprinting was compromised we



k

	Total col./ 100,000 cells	ES-like col./ total col. (%)	iPS line/ ES-like col. (%)	Estimated efficiency (%)
Oct4	156 ± 31	24.0 ± 7	22	0.080
Nanog	947 ± 187	11.5 ± 4	5	0.050

Figure 1 | Generation of Oct4- and Nanog-selected iPS cells. **a**, Targeting strategy to generate an Oct4-IRES-GFPneo allele. The resulting GFPneo fusion protein has sufficient neomycin-resistance activity in ES cells; GFP fluorescence, however, is not visible. **b**, Phase-contrast micrograph of Oct4 iPS cells (clone 18) grown on irradiated MEFs. Inset: an ES-cell-like colony 5 days after seeding in clonal density without feeder cells. iPS clone 18 cells exhibited strong alkaline phosphatase activity (**c**) and were homogenously labelled with antibodies against SSEA1 (**d**, **e**) and Nanog (**f**, **g**). **h**, One example of an ES-like colony 16 days after infection (left). Most G418-resistant colonies, however, consisted of flat non-ES-like cells (right): **b**, 10×; **c**–**g**, 20×; **h**, 4×. **i**, Gradual activation of the Nanog and Oct4-neo alleles. Shown are the total colony numbers of one experiment at day 20 after infection starting neomycin selection at day 3, 6 and 9. **j**, Fraction of total selected cells expressing alkaline phosphatase, SSEA1 and Nanog 0, 14, and 20 days after infection (counted were more than ten visual fields containing $n > 1{,}000$ total cells for every time point; error bars indicate s.d.). **k**, Estimated reprogramming efficiency of Oct4 selection and Nanog selection ($n = 3$ different experiments; s.e.m. is shown). Indicated are the total number of drug-resistant colonies per 100,000 plated MEFs 20 days after infection; the fraction of ES-like colonies per total number of colonies; the fraction of iPS cell lines that could be established from picked ES-like colonies as defined by homogenous alkaline phosphatase, SSEA1 and Nanog expression. After determining the fraction of *Sox2*- (83.4%), *Oct4*- (53.2%) and *c-myc*- (46.3%) infected MEFs 2 days after infection by immunofluorescence and assuming 50% were infected by *Klf4* viruses, we estimated the overall reprogramming efficiency as the ratio of quadruple-infected cells and the extrapolated total number of iPS cell lines that could be established with G418 selection starting at day 6 after infection.

assessed the methylation status of the four imprinted genes *H19*, *Peg1* (also called *Mest*), *Peg3* and *Snrpn*. As shown in Fig. 2e, bands corresponding to an unmethylated and methylated allele were detected for each gene in MEFs, iPS cells and TTFs. In contrast, embryonic germ cells, which have erased all imprints[8], were unmethylated. Our results indicate that the epigenetic state of the *Oct4* and *Nanog* genes was reprogrammed from a transcriptionally repressed (somatic) to an active (embryonic) state and that the pattern of somatic imprinting was maintained in iPS cells. Furthermore, the presence of imprints suggests a non-embryonic-germ-cell origin of iPS cells.

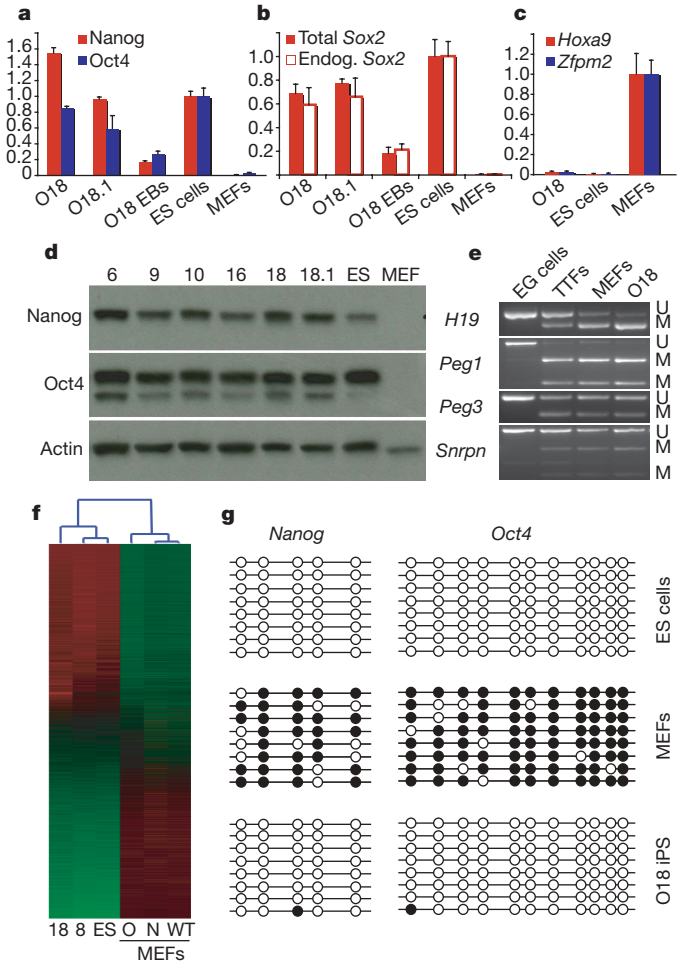


Figure 2 | Expression and promoter methylation analysis of iPS cells. **a**–**c**, qRT–PCR analysis ($n = 3$ independent PCR reactions; error bars indicate s.d.) of Oct4 iPS clone 18, subclone 18.1, 2-week-old embryoid bodies (EBs) derived from clone 18, V6.5 ES cells and Oct4-neo MEFs shows similar Nanog (red bars) and total Oct4 (blue bars) levels as in ES cells (**a**); slightly lower total Sox2 levels (filled red bars), mostly due to expression of endogenous Sox2 transcripts (open red bars, **b**); and strong downregulation of Hoxa9 (red) and Zfpm2 (blue) transcripts in iPS cells (**c**). Transcript levels were normalized to Gapdh expression, with expression levels in ES cells (**a**, **b**) and MEFs (**c**) set as 1. **d**, Western blot analysis for Oct4 and Nanog expression of different Oct4 iPS clones (6, 9, 10, 16, 18) and a GFP-labelled subclone of clone 18 (18.1). **e**, COBRA methylation analysis[32] of imprinted genes *H19* (maternally expressed), *Peg1* (paternally expressed), *Peg3* (paternally expressed) and *Snrpn* (paternally expressed). Upper band, unmethylated (U); lower band, methylated (M). **f**, Unsupervised hierarchical clustering of averaged global transcriptional profiles obtained from Oct4-neo iPS clone 18, Nanog-neo iPS clone 8, genetically matched ES cells (V6.5;129SvJae/C57Bl/6), Oct4-neo MEFs (O), Nanog-neo MEFs (N) and wild-type 129/B6 F1 MEFs (WT). **g**, Analysis of the methylation state of the *Oct4* and *Nanog* promoters using bisulphite sequencing. Open circles indicate unmethylated and filled circles methylated CpG dinucleotides. Shown are eight representative sequenced clones from ES cells (V6.5), Oct4-neo MEFs and Oct4-neo iPS clone 18.

Chromatin modifications

Recently, downstream target genes of *Oct4*, *Nanog* and *Sox2* have been defined in ES cells by genome-wide location analyses[9,10]. These targets include many important developmental regulators, a proportion of which is also bound and repressed by PcG (Polycomb-Group) complexes[11,12]. Notably, the chromatin at many of these non-expressed target genes adopts a bivalent conformation in ES cells, carrying both the 'active' histone H3 lysine 4 (H3K4) methylation mark and the 'repressive' histone H3 lysine 27 (H3K27) methylation mark[13,14]. In differentiated cells, those genes tend instead to carry either H3K4 or H3K27 methylation depending on their expression state. We used chromatin immunoprecipitation (ChIP) and real-time PCR to quantify H3K4 and H3K27 methylation for a set of

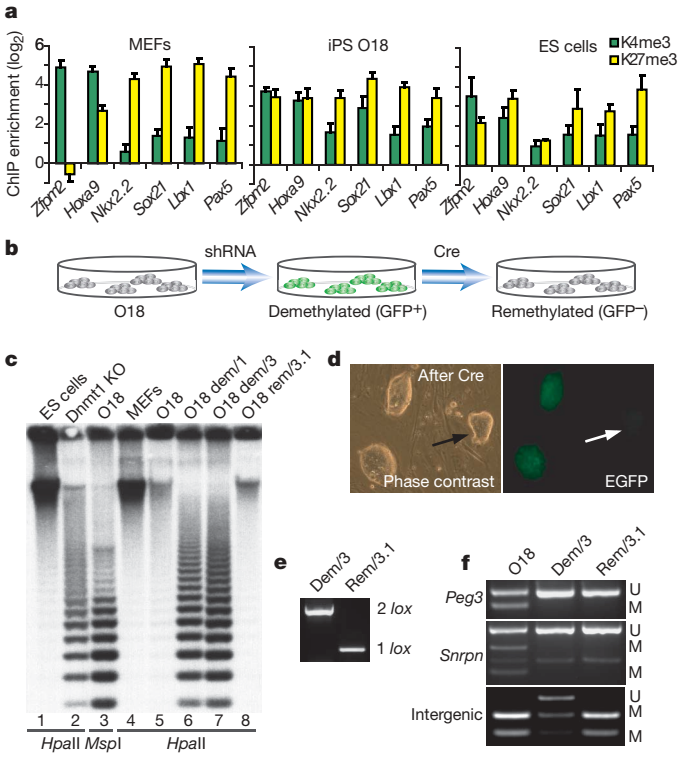


Figure 3 | Reprogrammed MEFs acquire an ES-cell-like epigenetic state. **a**, Real-time PCR after chromatin immunoprecipitation using antibodies against tri-methylated histone H3K4 and H3K27. Shown are the $\log_2$ enrichments for several previously reported 'bivalent' loci in ES cells ($n = 3$ experiments; error bars indicate s.d.). *Zfpm2* and *Hoxa9* show enrichment for the active (H3K4) mark in MEFs and are expressed (Fig. 1c and microarray data), whereas the other tested genes remain silent (microarray data). All loci tested in iPS clone O18 show enrichment for both H3K4 and H3K27 tri-methylation ('bivalent'), as seen in ES cells (V6.5). (See Supplementary Fig. 2 for H3K4 and H3K27 tri-methylation analysis of a subclone (clone O18.1) and Nanog-neo iPS clone N8.) **b**, Experimental design to de- and remethylate genomic DNA. Clone O18 was infected with the *Dnmt1*-hairpin-containing lentiviral vector pSicoR-GFP. The shRNA and GFP marker in the pSicoR vector are flanked by *loxP* sites[18]. Green colonies were expanded and passaged four times. Tat-Cre protein transduction was used to remove the shRNA[33]. **c**, Southern blot analysis of the minor satellite repeats using a methylation-sensitive restriction enzyme (*Hpa*II) and its methyl-insensitive isoschizomer (*Msp*I) as a control. Loss of methylation in two different clones (lanes 6 and 7) is comparable to Dnmt1 knockout ES cells (lane 2). After Cre-mediated recombination, complete remethylation (lane 8) of the repeats is observed within four passages. **d**, **e**, Successful loop out after Tat-Cre treatment was identified by disappearance of EGFP fluorescence (arrow) and verified by PCR analysis (**e**). **f**, COBRA assay of the imprinted genes *Peg3* and *Snrpn* and a random intergenic region close to the *Otx2* locus (Intergenic), demonstrating the expected resistance to *de novo* methylation of imprinted genes in contrast to non-imprinted intergenic sequences. U, unmethylated band; M, methylated band.

genes reported to be bivalent in pluripotent ES cells[13]. Figure 3a shows that the fibroblast-specific genes *Zfpm2* and *Hoxa9* carried stronger H3K4 methylation than H3K27 methylation in the donor MEFs, whereas the silent genes *Nkx2.2*, *Sox1*, *Lbx1* and *Pax5* primarily carried H3K27 methylation. In contrast, in the Oct4 iPS cells, all of these genes showed comparable enrichment for both histone modifications, similar to normal ES cells (Fig. 3a). Identical results were obtained in Nanog iPS clones selected from Nanog-neo MEFs (Supplementary Fig. 2). These data suggest that the chromatin configuration of somatic cells is re-set to one that is characteristic of ES cells.

iPS cells tolerate genomic demethylation

Tolerance of genomic demethylation is a unique property of ES cells in contrast to somatic cells, which undergo rapid apoptosis on loss of the DNA methyltransferase Dnmt1 (refs 15–17). We investigated whether iPS cells would be resistant to global demethylation after Dnmt1 inhibition and would be able to re-establish global methylation patterns after restoration of Dnmt1 activity. To this end, we used a conditional lentiviral vector harbouring a *Dnmt1*-targeting short hairpin (sh)RNA and a green fluorescence protein (GFP) reporter gene (Fig. 3b and ref. 18). Infected iPS cells were plated at low density and GFP-positive colonies were picked and expanded. Southern blot analysis using *Hpa*II-digested genomic DNA showed that global demethylation of infected iPS cells (Fig. 3c, lanes 6, 7) was similar to *Dnmt1*$^{-/-}$ ES cells (lane 2). In contrast, uninfected iPS cells or MEFs (lanes 4, 5) displayed normal methylation levels. Morphologically, the GFP-positive cells were indistinguishable from the parental line or from uninfected sister subclones, indicating that iPS cells tolerate global DNA demethylation. In a second step, the *Dnmt1* shRNA was excised through Cre-mediated recombination and GFP-negative clones were picked (Fig. 3d). The cells had excised the shRNA vector (Fig. 3e) and normal DNA methylation levels were restored (Fig. 3c, lane 8) and were able to generate chimaeras (see below, Table 1), as has been reported previously for ES cells[19]. These observations imply that the *de novo* methyltransferases Dnmt3a and Dnmt3b were reactivated in iPS cells[20], leading to restoration of global methylation levels. As expected[19], the imprinted genes *Snrpn* and *Peg3* were unmethylated and resistant to remethylation (Fig. 3f).

Maintenance of the pluripotent state

Southern blot analysis indicated that Oct4-neo iPS clone 18 carried four to six copies of the *Oct4*, c-*myc* and *Klf4* retroviral vectors and only one copy of the *Sox2* retroviral vector (Fig. 4a). Because these four factors were under the control of the constitutively expressed retroviral long terminal repeat, it was unclear in a previous study why iPS cells could be induced to differentiate[7]. To address this question, we designed primers specific for the four viral-encoded transcription factor transcripts and compared expression levels by qRT–PCR in MEFs 2 days after infection in iPS cells, in embryoid bodies derived from iPS cells, and in demethylated and remethylated iPS cells (Fig. 4b). Although the MEFs represented a heterogeneous population composed of uninfected and infected cells, virally encoded RNA levels of *Oct4*, *Sox2* and *Klf4* RNA were 5-fold higher and of *c-myc* more than 10-fold higher than in iPS cells. This suggests silencing of the viral long terminal repeat by *de novo* methylation during the reprogramming process. Accordingly, the total *Sox2* and *Oct4* RNA levels in iPS cells were similar to those in wild-type ES cells, and the *Sox2* transcripts in iPS cells were mostly, if not exclusively, transcribed from the endogenous gene (compare Fig. 2b). On differentiation to embryoid bodies, both viral and endogenous transcripts were downregulated. All viral *Sox2*, *Oct4* and *Klf4* transcripts were upregulated by approximately twofold in Dnmt1 knockdown iPS cells, and again downregulated on restoration of Dnmt1 activity. This is consistent with previous data that Moloney virus is efficiently *de novo* methylated and silenced in embryonic but not in somatic cells[21,22]. Transcript levels of *c-myc* were about 20-fold lower in iPS cells than in infected MEFs, and did not change on differentiation or demethylation.

To follow the kinetics of vector inactivation during the reprogramming process, we isolated RNA from drug-resistant cell populations at different times after infection. Figure 4c shows that the viral-vector-encoded transcripts were gradually silenced during the transition from MEFs to iPS cells with a time course that corresponded to the gradual appearance of pluripotency markers (compare Fig. 1j). Finally, to visualize directly Oct4 and Nanog expression during differentiation, we injected Oct4 iPS cells into SCID mice to induce teratoma formation (Fig. 4d). Immunostaining revealed that Oct4 and Nanog were expressed in the centrally located undifferentiated cells but were silenced in the differentiated parts of the teratoma (Fig. 4e, f). Our results suggest that the retroviral vectors are subject to gradual silencing by *de novo* methylation during the reprogramming process. The maintenance of the pluripotent state and induction of differentiation strictly depends on the expression and normal regulation of the endogenous *Oct4* and *Nanog* genes.

Developmental potency

We determined the developmental potential of iPS cells by teratoma and chimaera formation. Histological and immunohistochemical analysis of Oct4- or Nanog-iPS-cell-induced teratomas revealed that the cells had differentiated into cell types representing all three embryonic germ layers (Supplementary Figs 3 and 4). To assess more stringently their developmental potential, various iPS cell lines were injected into diploid (2N) or tetraploid (4N) blastocysts. After injection into 2N blastocysts both Nanog iPS and Oct4 iPS clones derived from MEFs (Fig. 5a) or from TTFs (Fig. 5b,c), as well as iPS cells that had been subjected to a consecutive cycle of demethylation and remethylation (compare Fig. 3b, c), efficiently generated viable

Table 1 | Summary of blastocyst infections

Cell line	2N injections				4N injections		
	Injected blastocysts	Live chimaeras	Chimaerism (%)	Germ line	Injected blastocysts	Dead embryos (arrested)	Live embryos (analysed)
O6	ND	ND	ND	ND	13	0	2 (E12.5)
O9	30	5	30–70	Yes	90	3 (E11–13.5)	12 (E10–12.5)
O16	15	3	10–30	Yes	ND	ND	ND
O18	95	8	5–50	No	134	7 (E9–11.5)	4* (E10–12.5)
O3-2	ND	ND	ND	ND	25	2 (E8,11.5)	0
O4-16	ND	ND	ND	ND	35	4 (E11–13.5)	3 (E14.5)
N7	30	1	30	ND	ND	ND	ND
N8	90	14	5–50	No	118	9 (E9–11.5)	1* (E12.5)
N14	30	5	5–20	ND	46	2 (E8,11.5)	1 (E12.5)
TT-O25	50	2	30†	ND	39	3 (E9.5)	0
O18 rem/3.1	25	1	30	ND	ND	ND	ND

The extent of chimaerism was estimated on the basis of coat colour or EGFP expression. ND, not determined. 4N injected blastocysts were analysed between embryonic day E10.5 and E14.5. 'Analysed' indicates the day of embryonic development analysed; 'arrested' indicates the estimated stage of development of dead embryos.
* Developmentally retarded or abnormal. O18 rem/3.1 is a de- and remethylated iPS clone (Fig. 3c).
† On the basis of GFP fluorescence.

high-contribution chimaeras (summarized in Table 1). To test for germline transmission, chimaeras derived from two different iPS lines (Oct4 iPS O9 and O16) were mated with normal females, and blastocysts were isolated and genotyped by three different PCR reactions for the presence of the multiple viral *Oct4* and *c-myc* genes and for the single-copy GFPneo sequences inserted into the *Oct4* locus of the donor cell (Fig. 1a). Figure 5f shows that 9 out of 16 embryos from two chimaeras were positive for the viral copies. As expected, only half of the viral-positive blastocysts contained the GFPneo sequences (5 out of 9 embryos, Fig. 5f, left panel). When embryonic day (E)10 embryos derived from an Oct4 iPS line O16 chimaera were genotyped, three out of eight tested embryos were transgenic (Fig. 5f, right panel). Finally, we injected iPS cells into 4N blastocysts as this represents the most rigorous test for developmental potency, because the resulting embryos are composed only of the injected donor cells ('all ES embryo'). Figure 5d, e shows that both Oct4 and Nanog iPS cells could generate mid- and late-gestation 'all iPS embryos' (summarized in Table 1). These findings indicate that iPS cells can establish all lineages of the embryo and thus have a similar developmental potential as ES cells.

Discussion

The results presented here demonstrate that the four transcription factors Oct4, Sox2, c-myc and Klf4 can induce epigenetic reprogramming of a somatic genome to an embryonic pluripotent state. In contrast to selection for Fbx15 activation[7], fibroblasts that had reactivated the endogenous *Oct4* (Oct4-neo) or *Nanog* (Nanog-neo) loci grew independently of feeder cells, expressed normal Oct4, Nanog and Sox2 RNA and protein levels, were epigenetically identical to ES cells by a number of criteria, and were able to generate viable chimaeras, contribute to the germ line and generate viable late-gestation embryos after injection into tetraploid blastocysts. Transduction of the four factors generated significantly more drug-resistant cells from Nanog-neo than from Oct4-neo fibroblasts but a higher fraction of Oct4-selected cells had all the characteristics of pluripotent ES cells, suggesting that *Nanog* activation is a less stringent criterion for pluripotency than *Oct4* activation.

Our data suggest that the pluripotent state of Oct4 iPS and Nanog iPS cells is induced by the virally transduced factors but is largely maintained by the activity of the endogenous pluripotency factors including Oct4, Nanog and Sox2, because the viral-controlled transcripts, although expressed highly in MEFs, become mostly silenced in iPS cells. The total levels of Oct4, Nanog and Sox2 were similar in iPS and wild-type ES cells. Consistent with the conclusion that the pluripotent state is maintained by the endogenous pluripotency genes is the finding that the *Oct4* and the *Nanog* genes became hypomethylated in iPS cells as in ES cells, and that the bivalent histone modifications of developmental regulators were re-established. Furthermore, iPS cells were resistant to global demethylation induced by inactivation of Dnmt1, similar to ES cells but in contrast to somatic cells. Re-expression of Dnmt1 in the hypomethylated ES cells resulted in global remethylation, indicating that the iPS cells had also reactivated the *de novo* methyltransferases Dnmt3a and Dnmt3b. All these observations are consistent with the conclusion that the iPS cells have gained an epigenetic state that is similar to that of normal ES cells. This conclusion is further supported by the recent observation that female iPS cells, similar to ES cells, reactivate the somatically silenced X chromosome[23].

Expression of the four transcription factors proved to be a robust method to induce reprogramming of somatic cells to a pluripotent state. However, the use of retrovirus-transduced oncogenes represents a serious barrier to the eventual use of reprogrammed cells for therapeutic application. Much work is needed to understand the molecular pathways of reprogramming and to eventually find small

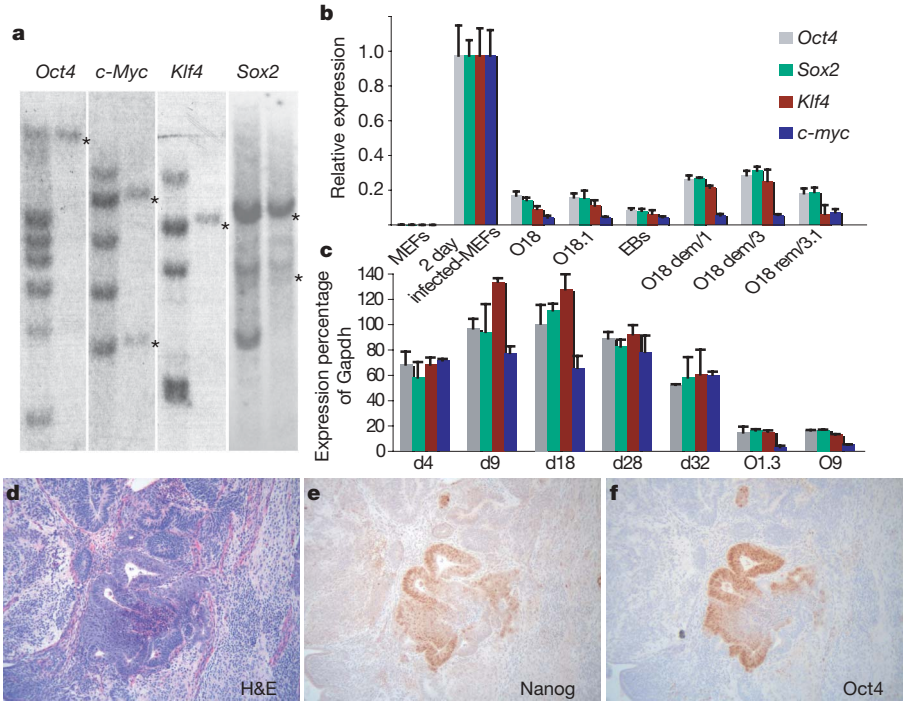


Figure 4 | Efficient silencing of retroviral transcripts in induced pluripotent cells. a, Southern blot analysis of proviral integrations in iPS clone O18 (left lanes) for the four retroviral vectors. Uninfected ES cells (right lanes) show only one or two bands corresponding to the endogenous gene (marked by an asterisk). **b**, Quantitative RT–PCR using primers specifically detecting the four viral transcripts. Shown are Oct4-neo iPS clone 18 and a GFP-labelled subclone, Oct4-neo MEFs, 2-week-old embryoid bodies generated from clone 18, two demethylated clones (18 dem/1 and 18 dem/3), a remethylated clone (18 rem/3.1), and Oct4-neo MEFs 2 days after infection with all four viruses but not selected with G418 ($n = 3$ independent experiments; error bars indicate s.d.). **c**, Viral transcript levels at various time points in cell populations after infection and Oct4 selection and in the two Oct4 iPS cell lines O1.3 and O9 ($n = 3$ independent experiments; error bars indicate s.d.). **d–f**, Paraffin sections of a teratoma 26 days after subcutaneous injection of Oct4 iPS clone 18 cells into SCID mice. H&E, haematoxylin and eosin. Nanog (**e**) and Oct4 (**f**) expression was confined to undifferentiated cell types as indicated an immunohistochemical analysis.

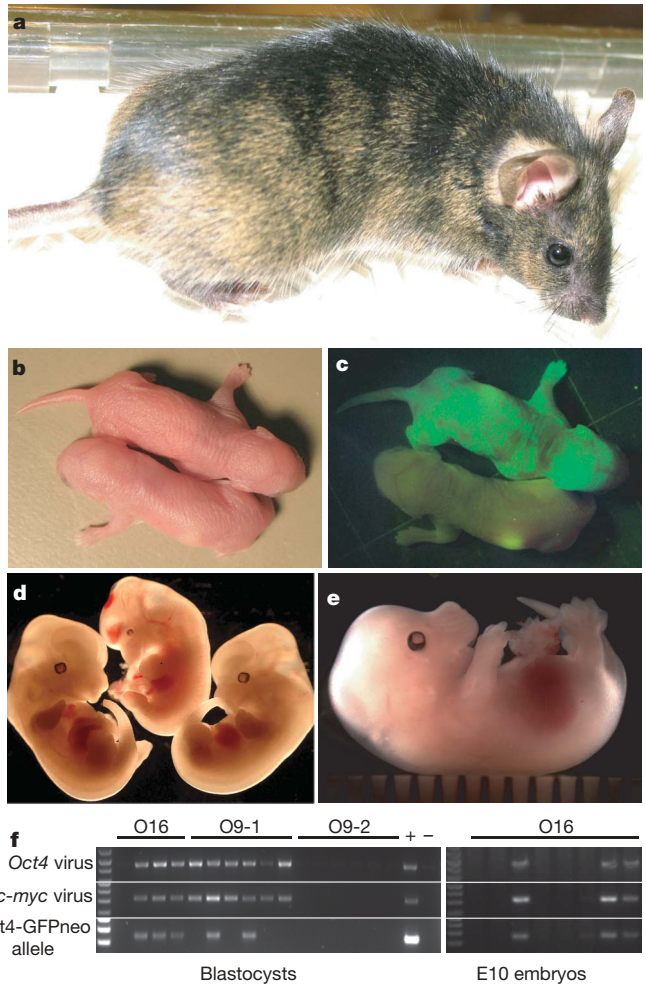


Figure 5 | Developmental pluripotency of reprogrammed fibroblasts. a, A 6-week-old chimaeric mouse. Agouti-coloured hairs originated from Oct4 iPS cell line O18.1. **b**, **c**, Two live pups after 2N blastocyst injection, one of which shows a high contribution (**c**) of the TTF-derived Oct4 iPS cell line TT-O25, which had been GFP-labelled with a lentiviral ubiquitin-EGFP vector. **d**, 'All iPS cell embryos' were generated by injection of iPS cells into 4N blastocysts[34]. Live E12.5 embryos generated from Oct4 iPS line O6 (left), from Nanog iPS line N14 (middle) and from V.6.5 ES cells (right) are shown. **e**, A normally developed E14.5 embryo was derived from Oct4 iPS cell line O4-16 after tetraploid complementation and was isolated by screening MEFs for activation of GFP inserted into the *Oct4* locus. **f**, Germline contribution of Oct4 iPS clones O9 and O16. Genotyping of blastocysts from females mated with three chimaeric males demonstrated the presence of *Oct4* and *c-myc* virus integrations and the Oct4-IRES-GFPneo allele (left panel). Because of the multiple integrations (Fig. 4a) all embryos with iPS cell contribution are expected to be positive for proviral sequences in this assay. In contrast, the single-copy Oct4-IRES-GFPneo allele segregated into only 5 of the 9 virus-positive embryos. All six blastocysts from O9 chimaera 1 were iPS-cell-derived, suggesting that this chimaera was a pseudo-male. Additional genotyping identified 13 out of 72 tested blastocysts derived from iPS line O9 and 4 out of 13 blastocysts derived from iPS line O16 chimaeras carrying the viral transgenes. The right panel shows that 3 out of 8 tested E.10 mid-gestation embryos were sired by a chimaera derived from the donor iPS line O16. +, positive control; −, negative control.

molecules that could achieve reprogramming without gene transfer of potentially harmful genes.

METHODS SUMMARY

Cell culture, gene targeting and viral infections. ES and iPS cells were cultivated on irradiated MEFs. Using homologous recombination we generated ES cells carrying an IRES-GFPneo fusion cassette downstream of *Oct4* exon 5 (Fig. 1a). The *Nanog* gene was targeted as described[24]. Transgenic MEFs were isolated and selected from E13.5 chimaeric embryos after blastocyst injection of Oct4-IRES-GFPneo- or Nanog-neo-targeted ES cells. MEFs were infected overnight with the Moloney-based retroviral vector pLIB (Clontech) containing the murine complementary DNAs of *Oct4*, *Sox2*, *Klf4* and *c-myc*.

Southern blot, methylation and chromatin analyses. To assess the levels of DNA methylation, genomic DNA was digested with *Hpa*II and hybridized to a probe for the minor satellite repeats[25] or with an IAP probe[26]. Bisulphite treatment was performed with the Qiagen EpiTect Kit. For the methylation status of *Oct4* and *Nanog* promoters, bisulphite sequencing analysis was performed as described previously[27]. For imprinted genes, a COBRA assay was performed. PCR primers and conditions were as described previously[28]. The status of bivalent domains was determined by chromatin immunoprecipitation followed by quantitative PCR analysis, as described previously[12]

Expression analysis. Total RNA was reverse-transcribed and quantified using the QuantTtect SYBR green RT–PCR Kit (Qiagen) on a 7000 ABI detection system. Western blot and immunofluorescence analysis was performed as described[29,30]. Microarray targets from 2 μg total RNA were synthesized and labelled using the Low RNA Input Linear Amp Kit (Agilent), hybridized to Agilent whole-mouse genome oligonucleotide arrays (G4122F) and analysed as previously described[31].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 February; accepted 22 May 2007.
Published online 6 June 2007.

1. Hochedlinger, K. & Jaenisch, R. Nuclear transplantation, embryonic stem cells, and the potential for cell therapy. *N. Engl. J. Med.* **349,** 275–286 (2003).
2. Yang, X. *et al.* Nuclear reprogramming of cloned embryos and its implications for therapeutic cloning. *Nature Genet.* **39,** 295–302 (2007).
3. Hochedlinger, K. & Jaenisch, R. Nuclear reprogramming and pluripotency. *Nature* **441,** 1061–1067 (2006).
4. Tada, M., Takahama, Y., Abe, K., Nakatsuji, N. & Tada, T. Nuclear reprogramming of somatic cells by *in vitro* hybridization with ES cells. *Curr. Biol.* **11,** 1553–1558 (2001).
5. Cowan, C. A., Atienza, J., Melton, D. A. & Eggan, K. Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. *Science* **309,** 1369–1373 (2005).
6. Jaenisch, R. Human cloning—the science and ethics of nuclear transplantation. *N. Engl. J. Med.* **351,** 2787–2791 (2004).
7. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
8. Labosky, P. A., Barlow, D. P. & Hogan, B. L. Mouse embryonic germ (EG) cell lines: transmission through the germline and differences in the methylation imprint of insulin-like growth factor 2 receptor (*Igf2r*) gene compared with embryonic stem (ES) cell lines. *Development* **120,** 3197–3204 (1994).
9. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
10. Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.* **38,** 431–440 (2006).
11. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125,** 301–313 (2006).
12. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441,** 349–353 (2006).
13. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
14. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature Cell Biol.* **8,** 532–538 (2006).
15. Jackson-Grusby, L. *et al.* Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genet.* **27,** 31–39 (2001).
16. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69,** 915–926 (1992).
17. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33,** 5868–5877 (2005).
18. Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proc. Natl Acad. Sci. USA* **101,** 10380–10385 (2004).
19. Holm, T. M. *et al.* Global loss of imprinting leads to widespread tumorigenesis in adult mice. *Cancer Cell* **8,** 275–285 (2005).
20. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99,** 247–257 (1999).
21. Stewart, C. L., Stuhlmann, H., Jähner, D. & Jaenisch, R. *De novo* methylation, expression, and infectivity of retroviral genomes introduced into embryonal carcinoma cells. *Proc. Natl Acad. Sci. USA* **79,** 4098–4102 (1982).
22. Jähner, D. *et al. De novo* methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* **298,** 623–628 (1982).
23. Maherali, N. *et al.* Global epigenetic remodeling in directly reprogrammed fibroblasts. *Cell Stem Cells* (in the press)..

24. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113,** 631–642 (2003).
25. Chapman, V., Forrester, L., Sanford, J., Hastie, N. & Rossant, J. Cell lineage specific undermethylation of mouse repetitive DNA. *Nature* **307,** 284–286 (1984).
26. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.* **20,** 116–117 (1998).
27. Blelloch, R. *et al.* Reprogramming efficiency following somatic cell nuclear transfer is influenced by the differentiation and methylation state of the donor nucleus. *Stem Cells* **24,** 2007–2013 (2006).
28. Lucifero, D., Mertineit, C., Clarke, H. J., Bestor, T. H. & Trasler, J. M. Methylation dynamics of imprinted genes in mouse germ cells. *Genomics* **79,** 530–538 (2002).
29. Hochedlinger, K., Yamada, Y., Beard, C. & Jaenisch, R. Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell* **121,** 465–477 (2005).
30. Wernig, M. *et al.* Functional integration of embryonic stem cell-derived neurons *in vivo*. *J. Neurosci.* **24,** 5258–5268 (2004).
31. Brambrink, T., Hochedlinger, K., Bell, G. & Jaenisch, R. ES cells derived from cloned and fertilized blastocysts are transcriptionally and functionally indistinguishable. *Proc. Natl Acad. Sci. USA* **103,** 933–938 (2006).
32. Eads, C. A. & Laird, P. W. Combined bisulfite restriction analysis (COBRA). *Methods Mol. Biol.* **200,** 71–85 (2002).
33. Peitz, M., Pfannkuche, K., Rajewsky, K. & Edenhofer, F. Ability of the hydrophobic FGF and basic TAT peptides to promote cellular uptake of recombinant Cre recombinase: a tool for efficient genetic engineering of mammalian genomes. *Proc. Natl Acad. Sci. USA* **99,** 4489–4494 (2002).
34. Eggan, K. *et al.* Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation. *Proc. Natl Acad. Sci. USA* **98,** 6209–6214 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Suh, D. Fu and J. Dausman for technical assistance; J. Love for help with the microarray analysis; S. Markoulaki for help with blastocyst injections; F. Edenhofer for a gift of Tat-Cre; and S. Yamanaka for the Nanog-neo construct. We acknowledge L. Zagachin in the MGH Nucleic Acid Quantitation core for assistance with real-time PCR. We also thank C. Lengner, C. Beard and M. Creyghton for constructive criticism. M.W. was supported in part by fellowships from the Human Frontiers Science Organization Program and the Ellison Foundation; B.B. by grants from the Burroughs Wellcome Fund, the Harvard Stem Cell Institute and the NIH; and R.J. by grants from the NIH.

Author Contributions M.W., A.M. and R.J. conceived and designed the experiments and wrote the manuscript; M.W. derived all iPS lines; M.W. and A.M. performed the *in vitro* and *in vivo* characterization of the iPS lines (teratoma, 2N and 4N injections and IHC) and the conditional Dnmt1 experiment; A.M. investigated the promoter and imprinting methylation; M.K. and B.B. performed and analysed the real-time PCRs and ChIP experiments; R.F. and K.H. generated the selectable MEFs and TTFs; R.F. performed western blot and PCR analyses; and T.B. performed the microarray analysis and the proviral integration Southern blots.

Author Information All microarray data from this study are available from Array Express at the EBI (http://www.ebi.ac.uk/arrayexpress) under the accession number E-MEXP-1037. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.J. (jaenisch@wi.mit.edu).

METHODS

Cell culture, MEF isolation, gene targeting and viral infections. ES and iPS cells were cultivated on irradiated MEFs in DME containing 15% fetal calf serum, leukaemia inhibiting factor (LIF), penicillin/streptomycin, L-glutamine, and non-essential amino acids. All cells were depleted of feeder cells for two passages on 0.2% gelatin before RNA, DNA or protein isolation. Transgenic MEFs were isolated and selected in 2 μg ml^{-1} puromycin (Sigma) from E13.5 chimaeric embryos after blastocyst injection of Oct4-inducible KH2 ES cells[29] that had been previously targeted with either Oct4-IRES-GFPneo or Nanog-neo constructs (Fig. 1a and ref. 24). Using homologous recombination in ES cells, an IRES-GFPneo fusion cassette was inserted into the *Bcl*I site downstream of *Oct4* exon 5. Correctly targeted ES cell clones were screened by Southern analysis of *Nco*I-digested DNA using a 5′ external probe. The murine cDNAs for *Oct4*, *Sox2*, *Klf4* and *c-myc* were PCR amplified from ES cell cDNA, sequence-verified and cloned into the Moloney-based retroviral vector pLIB (Clontech). 2×10^5 MEFs or TTFs at passage 2–4 were infected overnight with pooled viral supernatant generated by transfection of 4×10^6 HEK293T cells (Fugene, Roche) with 10 μg of viral vectors and the packaging plasmid pCL-Eco in a 10-cm dish[35].

Blastocyst injection. Diploid or tetraploid blastocysts (94–98 h after HCG injection) were placed in a drop of DMEM with 15% FCS under mineral oil. A flat-tip microinjection pipette with an internal diameter of 12–15 mm was used for iPS cell injection (using a Piezo micromanipulator[34]). A controlled number of cells was injected into the blastocyst cavity. After injection, blastocysts were returned to KSOM media and placed at 37 °C until transferred to recipient females.

Recipient females and caesarean sections. Ten to fifteen injected blastocysts were transferred to each uterine horn of 2.5 days post coitum pseudo-pregnant B6D2F1 females. To recover full-term pups, recipient mothers were killed at 19.5 days post coitum. Surviving pups were fostered to lactating BALB/c mothers.

Southern blot, methylation and chromatin analyses. To assess the levels of DNA methylation, genomic DNA was digested with *Hpa*II, and hybridized to pMR150 as a probe for the minor satellite repeats[25], or with an IAP-probe[26]. Bisulphite treatment was performed with the Qiagen EpiTect Kit. For the methylation status of *Oct4* and *Nanog* promoters, bisulphite sequencing analysis was performed as described previously[27]. A total of 10–20 clones of each sample was sequenced in both directions. For imprinted genes, a COBRA assay was performed. PCR primers and conditions were as described previously[28]. PCR products after bisulphite treatment and gel purification were digested with *Bst*UI (CGCG; *H19*, *Peg3* and *Snrpn*) or HpyCH4 IV (ACGT; *Peg1*) and resolved on a 2% agarose gel. Unmethylated CpGs in the recognition sequence will be converted to T and not cut. The status of bivalent domains was determined by chromatin immunoprecipitation followed by quantitative PCR analysis, as described previously[12].

Expression analysis. Fifty nanograms of total RNA isolated using TRIzol reagent (Invitrogen) was reverse-transcribed and quantified using the QuantTtect SYBR green RT–PCR Kit (Qiagen) on a 7000 ABI detection system. Western blot and immunofluorescence analysis was performed as described[29,30]. Primary antibodies included Oct4 (monoclonal mouse, Santa Cruz), Nanog (polyclonal rabbit, Bethyl), actin (monoclonal mouse, Abcam) and SSEA1 (monoclonal mouse, Developmental Studies Hybridoma Bank). Fluorophor-labelled secondary antibodies were purchased from Jackson Immunoresearch. Microarray targets from 2 μg total RNA were synthesized and labelled using the Low RNA Input Linear Amp Kit (Agilent) and hybridized to Agilent whole-mouse genome oligo arrays (G4122F). Arrays were scanned on an Agilent G2565B scanner and signal intensities were calculated in Agilent FE software. Data sets were normalized using an R script (available at http://www.ebi.ac.uk/arrayexpress) and clustered as previously described[31].

Viral integrations. Genomic DNA was digested with *Spe*I (Oct4, c-myc, klf4) or HindIII (Sox2) overnight, followed by electrophoresis and transfer, and the blots were hybridized to the respective radioactively labelled cDNAs of the four transcription factors.

Genotyping. Blastocysts were lysed for 4 h in 10 μl 50 mM Tris, pH 8.8, containing 1 mM EDTA, 0.5% Tween20 and 200 μg ml^{-1} proteinase K. After heat inactivation for 15 min, PCR was performed with the following conditions: 95 °C 30 s (1 cycle); 95 °C 10 s, 60 °C 15 s, 72 °C 15 s (40 cycles); 72 °C 5 min.

Primer sequences for genotyping. GFP-F, TCCATGGCCAACACTAGTCA; GFP-R, TCCCAGAATGTTGCCATCTT; pLIB-FW1, CCCCCTTGAACCTCCTCGTTCGAC; Oct4R, GAGGTTCCCTCTGAGTTGCTTT; MycR, CGAATTTCTTCCAGATATCCTCAC.

Primer sequences for viral-specific qRT–PCR. rtKlf4_virusF1, TCTCTAGGCGCCGGAATTC; rtKlf4_virusR1, CCATGTCAGACTCGCCAGGT; rtMyc_virusF1, CTTCTCTAGGCGCCGGAATT; rtMyc_virusR1, TGGTGAAGTTCACGTTGAGGG; rtOct4_virusF1, TACACCCTAAGCCTCCGCCT; rtOct4_virusR1, ATTCCGGCGCCTAGAGAAG; rtSox2_virusF1, TACACCCTAAGCCTCCGCCT; rtSox2_virusR1, ATTCCGGCGCCTAGAGAAG.

***Dnmt1* hairpin target sequence DZ.** GGAAAGAGATGGCTTAACA.

35. Naviaux, R. K., Costanzi, E., Haas, M. & Verma, I. M. The pCL vector system: rapid production of helper-free, high-titer, recombinant retroviruses. *J. Virol.* **70**, 5701–5705 (1996).

ARTICLES

Conformational entropy in molecular recognition by proteins

Kendra King Frederick[1], Michael S. Marlow[1], Kathleen G. Valentine[1] & A. Joshua Wand[1]

Molecular recognition by proteins is fundamental to almost every biological process, particularly the protein associations underlying cellular signal transduction. Understanding the basis for protein–protein interactions requires the full characterization of the thermodynamics of their association. Historically it has been virtually impossible to experimentally estimate changes in protein conformational entropy, a potentially important component of the free energy of protein association. However, nuclear magnetic resonance spectroscopy has emerged as a powerful tool for characterizing the dynamics of proteins. Here we employ changes in conformational dynamics as a proxy for corresponding changes in conformational entropy. We find that the change in internal dynamics of the protein calmodulin varies significantly on binding a variety of target domains. Surprisingly, the apparent change in the corresponding conformational entropy is linearly related to the change in the overall binding entropy. This indicates that changes in protein conformational entropy can contribute significantly to the free energy of protein–ligand association.

Numerous structural studies have revealed that protein–protein interfaces often involve dozens of amino acid residues and thousands of Å^2 of contact area[1]. It has also become apparent that a non-uniform contribution of individual residues to the free energy of binding can exist and that static structural analyses can mask important factors underlying the high-affinity interactions between proteins[2]. Of particular interest here is the role of protein conformational entropy in modulating the free energy of the association of a protein with a ligand. A simplistic decomposition emphasizes the fact that the entropy of binding (ΔS_{bind}), obtainable by calorimetric methods, is comprised of contributions associated with the protein, the ligand and the solvent:

$$\Delta G_{\mathrm{bind}} = \Delta H_{\mathrm{bind}} - T\Delta S_{\mathrm{bind}} = \Delta H_{\mathrm{bind}} - T(\Delta S_{\mathrm{protein}} + \Delta S_{\mathrm{ligand}} + \Delta S_{\mathrm{solvent}}) \quad (1)$$

It is well established that the transitions of a ligand from a disordered (high entropy) unbound state to a structured (lower entropy) bound state can profoundly influence the entropy of macromolecular associations[3]. It is also well established that burial of hydrophobic surface area and the consequent release of hydration waters to the bulk solvent can also contribute significantly to the thermodynamics of binding[4]. What is less understood is the potential entropic contributions from a 'structured' protein ($\Delta S_{\mathrm{protein}}$), which includes changes in its conformational entropy (ΔS_{conf}) as well as changes in rotational and translational entropy[5–7]. Here we focus on ΔS_{conf}. As is obvious from equation (1), the measurement of total system thermodynamic parameters does not resolve contributions from internal protein conformational entropy. The estimation of changes in conformational entropy due to protein–protein association from molecular dynamics simulations remains a considerable challenge[8]. Experimental measurement of the conformational entropy of the protein in its free and complexed states is therefore required. Recent developments in nuclear magnetic resonance (NMR) relaxation methods and analysis now make this feasible.

The conformational entropy of proteins is manifested as motion between different structural states[7]. This opens the door to using motion as a proxy for conformational disorder or entropy. In principle, the measurement of a protein's internal dynamics should facilitate characterization of conformational entropy through a 'counting of states' implicit in molecular motion[9]. Solution NMR spectroscopy is particularly well suited to measuring conformational dynamics over a wide-range of time scales[10]. Simple considerations lead to the conclusion that the motion expressed on the sub-nanosecond timescale corresponds to significant conformational entropy[7,9]. This timescale is directly accessed using NMR relaxation methods[9].

Calmodulin as a model system

We have employed calmodulin as a model system to investigate the role for changes in protein conformational entropy in the high-affinity association of proteins. Calmodulin is a central participant in the calcium-mediated signal transduction pathways of eukaryotes[11]. It interacts with and regulates the activity of approximately three-hundred proteins[12]. Previously, using NMR relaxation methods, we have shown that calcium-saturated calmodulin (CaM) is an unusually dynamic protein and is characterized by a broad, non-uniform multi-modal distribution of the amplitude of fast side-chain dynamics[13]. Binding of a target domain to CaM causes a significant redistribution of the fast side-chain dynamics in calmodulin[13]. This raises the possibility that CaM employs its internal conformational entropy to 'tune' its affinity for ligands.

Here we explore this question by using NMR methods to determine the dynamic response of human CaM (GenBank AAD45181) to the binding of six peptides representing the calmodulin-binding domains of the smooth muscle myosin light chain kinase (smMLCK; AAA69964)[14], the neuronal and endothelial nitric oxide synthases (nNOS and eNOS; AAB60654 and AAH63294, respectively)[15], the calmodulin kinase kinase (CaMKKα; EDM05132)[16], the calmodulin kinase I (CaMKI; EAW63990)[17] and the phosphodiesterase (PDE; AAD40738)[18]. Here we will use the nomenclature smMLCK(p) to emphasize the fact that we are employing peptide models of the calmodulin-binding domains of the regulated proteins. All of the calmodulin-binding domain peptides have a basic amphiphillic character and form α-helical structure when bound to calmodulin (Supplementary Table 1). Four of the peptides have been found previously

[1]Johnson Research Foundation and Department of Biochemistry & Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

by isothermal titration calorimetry to have roughly the same affinity for calmodulin but with widely different thermodynamic parameters defining the free energy of association[19,20]. We have repeated the isothermal titration calorimetry measurements at a temperature (35 °C) that is more optimal for solution NMR spectroscopy and have characterized the thermodynamics of binding of two additional domains (Fig. 1). In the case of the CaMKKα(p) and smMLCK(p) domains, binding is driven by a large favourable change in total binding enthalpy overcoming a large unfavourable change in total binding entropy. At the other extreme, nNOS(p) binding is driven by a favourable change in total enthalpy accompanied by a small favourable change in entropy. The PDE(p), CaMKI(p) and eNOS(p) calmodulin-binding domains represent intermediate cases. The entropy of binding of these domains varies by 90 kJ mol^{-1} and changes sign (Fig. 1).

Titration of CaM with each of the peptides reveals that all six of the resulting complexes have a 1:1 stoichiometry and are in slow exchange with their dissociated components on the NMR ^{1}H chemical shift timescale (not shown). The CaM–smMLCK(p), CaM–PDE(p) and CaM–CaMKKα(p) complexes have very little conformational heterogeneity, as judged by ^{15}N- and ^{13}C-heteronuclear single quantum correlation (HSQC) spectra, whereas the CaM–nNOS(p), CaM–eNOS(p) and CaM–CaMKI(p) complexes show some heterogeneity at a small number of locations in the calmodulin molecule. This was found to largely arise from populations of minor rotameric orientations of methyl-bearing side chains. These results indicate a range of localized conformational heterogeneity in calmodulin across the six calmodulin complexes. This heterogeneity represents classical conformational entropy.

The dynamic response of calmodulin

The sub-nanosecond (sub-ns) dynamics of the polypeptide backbone of calmodulin in the six complexes were probed using NMR relaxation techniques. The degree of spatial restriction of each motional probe was assigned a number between 0, corresponding to complete isotropic disorder, and 1, corresponding to a fixed orientation in the molecular frame. This parameter is the squared generalized order parameter (O^2) as it applies to the amide N-H bond (O^2_{NH}), the C_α-C' bond ($O^2_{C\alpha CO}$) and the methyl symmetry axis (O^2_{axis}). O^2_{NH} parameters at amide nitrogen sites were obtained from measurements of ^{15}N dipolar relaxation[21]. $O^2_{C\alpha CO}$ parameters were obtained from measurement of transverse cross-correlated relaxation between ^{13}CO chemical shift anisotropy and the ^{13}CO–^{13}Cα dipolar interactions[22]. The motion of methyl groups (O^2_{axis}) of calmodulin side chains were characterized using ^{2}H spin relaxation methods[23].

The dynamics of the backbone of calmodulin are invariant across the complexes, as indicated by the average O^2_{NH} and $O^2_{C\alpha CO}$ parameters (Supplementary Table 3). In contrast, the motion of methyl-bearing amino acid side chains varies significantly with the nature of the target domain. There are 56 methyl-bearing amino acids providing 80 methyl groups as probes distributed across the primary sequence of calmodulin and including 9 methionines that line the target domain binding sites formed in the various complexes.

Dynamics as a proxy for entropy

We are guided by Karplus and co-workers[7] to connect the change in internal protein dynamics to the conformational entropy, describing the protein as a disjoint multidimensional harmonic well:

$$S_{conf} = \Sigma p_i S^h{}_i - k_B \Sigma p_i \ln p_i \quad (2)$$

where $S^h{}_i$ represents the entropy manifested by fast intra-well motion and the second term corresponds to the classical conformational entropy arising from the $i = 1 \ldots N$ distinct conformations. Here $S^h{}_i$ is obtained from interpretation of local order parameters, which is model-dependent. Our approach finds its modern roots in the work of Akke *et al.*[24], in which a specific motional model (potential energy function) is used to provide a parametric relationship between what is measured, the squared generalized order parameter, and what is sought, a thermodynamic quantity such as the entropy (see Supplementary Fig. 1). We choose a simple harmonic oscillator treatment to make this connection[25]. It is important to point out that the absolute entropies obtained in this way are very dependent on the details of the potential energy function but that differences in entropy calculated from changes in O^2 are fairly insensitive to the model used[25,26]. As the reference state for obtaining ΔS_{conf} we use calcium-saturated calmodulin. The second term of equation (2) represents classical entropy arising from the local heterogeneity of side-chain conformers. This can be manifested on a range of timescales. Some methyl sites exhibited slowly interconverting conformational heterogeneity on the chemical shift timescale. This was interpreted as classical entropy with the population of each state (p_i) estimated from the intensity of cross peaks. This contributed less than 2% of the estimated change in conformational entropy due to binding. It has also been shown that fast motion between rotamer wells contributes significantly to low O^2_{axis} parameters[26]. This also represents conformational entropy and was estimated using a previously described model[26]. This resulted in a roughly constant 15% of the total conformational entropy. Further details of the calculation are provided in Supplementary Table 4.

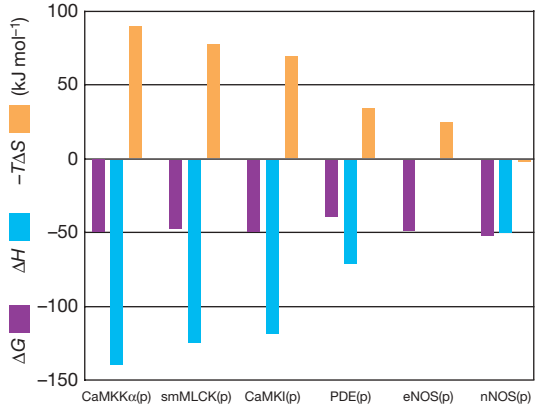


Figure 1 | Thermodynamic origins of high-affinity binding of target domains by calmodulin. The amino acid sequences of the domains are provided in Supplementary Table 1. Shown are the Gibbs free energy (ΔG), enthalpy (ΔH) and entropy ($-T\Delta S$) for the formation of the six calcium-saturated CaM-peptide complexes at 35 °C, as determined by isothermal titration calorimetry. Values are tabulated in Supplementary Table 2.

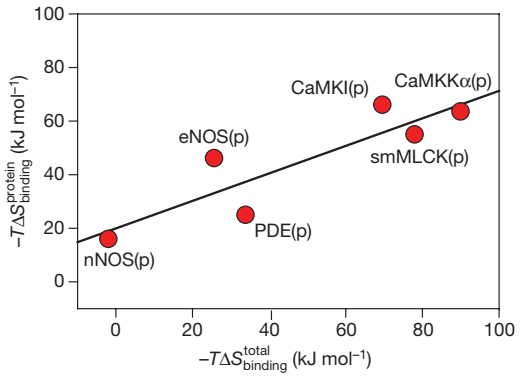


Figure 2 | Correlation of the change in conformational entropy of calmodulin with the change in the total entropy of binding of a target domain. The change in conformational entropy was estimated using equation (2), as described in Methods and Supplementary Information. Propagation of measurement error in fitted order parameters results in uncertainties in conformational entropy less than the size of the symbols used. The fitted linear correlation coefficient (R^2) of conformational entropy versus the entropy of binding is 0.78 with a slope of 0.51.

The conformational entropy of binding

The simple and direct interpretation of changes in dynamics as changes in conformational entropy is model-dependent and is therefore somewhat sensitive to the underlying accuracy of the model used. In addition, the presence of correlated motion in the packed protein interior will tend to result in an overestimate when interpreting each dynamic probe as independent (that is, by simple summation, equation (2)). Future work will undoubtedly refine the basic approach. Notwithstanding these limitations, the changes in the conformational entropy of calmodulin on binding to the six peptides, obtained by simple summation of the individual local entropies, shows a remarkable linear correlation ($R^2 = 0.78$) with the corresponding entropy of binding (Fig. 2). Taken at face value, half of the binding entropy is reflected in the motion of the methyl-bearing amino acid side chains. There is no a priori reason for such a correlation. However, the linearity of the correlation implies that either the change in the conformational entropy of calmodulin on binding a target domain is a major contribution to the binding entropy or that the various sources of entropy change in concert (see equation (1)). Regardless, it seems that the conformational entropy of calmodulin can vary sufficiently to impact the free energy changes arising from high-affinity protein associations. This model-dependent interpretation of the entropic significance of the observed changes in dynamics across the calmodulin complexes is buttressed by a relatively model-independent analysis described below.

The binding of smMLCK(p) to CaM results in a distribution of O^2_{axis} parameters that is remarkable for its distinct clustering into three apparent classes of motion[27]. The sum of the distributions of methyl group O^2_{axis} parameters in the six calmodulin complexes is shown in Fig. 3a. The large number of samplings ($n = 404$) provides for robust fitting of the distribution to the sum of three gaussians (see Supplementary Information). The best-fitted line is shown ($R^2 = 0.94$ and $P < 0.0001$) and the nine best-fitted parameters are provided in Supplementary Table 5. The summed

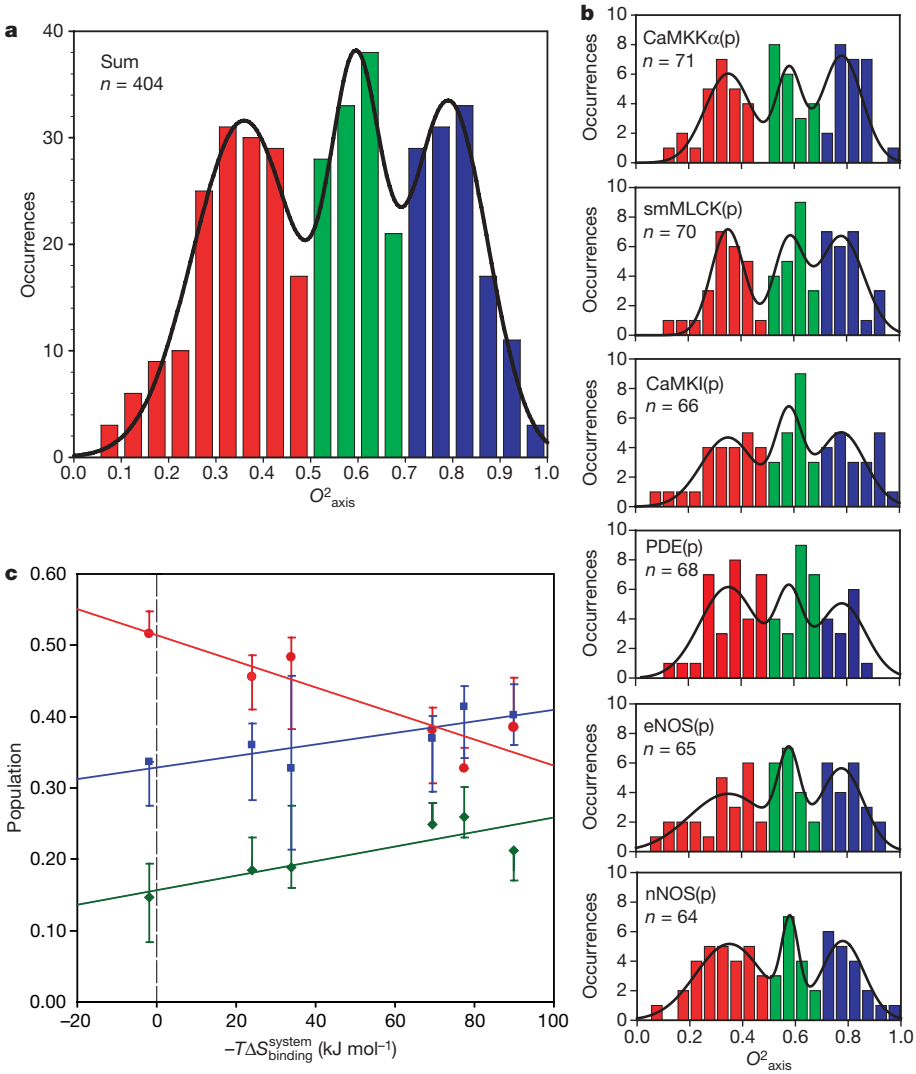


Figure 3 | Distribution of the amplitude of methyl-bearing side-chain motion of calmodulin in complex with target domains, and correlation with the change in total entropy of binding. a, Histogram of the sum of the O^2_{axis} parameter distributions of calmodulin in the six individual complexes obtained at 35 °C. The solid line represents the best-fitted solution to a 3-gaussian distribution with all nine parameters fitted. The best-fitted parameters are given in Supplementary Table 5. **b**, Histograms of the O^2_{axis} parameter distributions of calmodulin in the individual complexes. The solid lines represent fitted 3-gaussian distributions centred on O^2_{axis} values of 0.35 (J-class, red), 0.58 (α-class, green) and 0.78 (ω-class, blue). The relative populations of each class were derived from the fitted 3-gaussian distributions for each complex. **c**, Correlation of the change in population of the J, α and ω classes with the $-T\Delta S_{bind}$ have fitted linear correlation coefficients (R^2) of −0.83, +0.74 and +0.71, respectively. Correlation of the number of sites assigned to each class by simple binning, as colour-coded, yielded similar results (see Supplementary Table 7). Error bars reflect the variation of the population of each motional class that results from an increase or decrease in the measured O^2_{axis} values by two standard deviations.

distribution yielded fitted 3-gaussian distributions centred on O^2_{axis} values of 0.35 (large 'amplitude' motion), 0.58 (intermediate 'amplitude' motion) and 0.78 (highly restricted motion). Using these centres, the distributions of O^2_{axis} parameters in each of the six physiologically relevant calmodulin complexes are also satisfactorily described by a sum of three gaussians (Fig. 3). The relative populations of these motional classes in calmodulin vary considerably across the six complexes.

Although the distinctive grouping of order parameters, seen across all six complexes studied here, is often obscured in other proteins[28], the motional origin of these classes is clear. In the case of calmodulin, two fundamental types of motion occurring on the sub-ns timescale are involved: motion within a rotamer well, and motion between rotamer wells of side-chain torsion angles[26]. It has been shown that the class of large amplitude motion centred on a O^2_{axis} value of ~0.35 generally involves a significant contribution from rotameric interconversion on the nanosecond or faster timescale because it leads to a significant averaging of scalar coupling (J) constants[26]. More recent experimental results[29] and theoretical simulations[30] suggests this to be general. The class of motion at the other extreme is centred on an O^2_{axis} value of ~0.8, which represents highly restricted motion within a rotamer well. The class of moderate motion centred on an O^2_{axis} value of ~0.6 involves little detectable rotamer interconversion and is restricted to motion within a single rotamer well. The precise value reflects intra-well motion and the effects of superposition of similar motion about connected torsion angles. We have termed these groupings the J-, ω- and α-classes of motion, respectively[9].

The fractional populations of each motional class, derived from the fitting of the observed distributions of O^2_{axis} parameters, in the six complexes reveal a surprising correlation with the change in total system entropy for binding (Fig. 3c). The population of the J-class is negatively correlated with the entropic contribution ($-T\Delta S$) to the free energy of binding. The populations of the ω- and α-classes are positively correlated. The correlations are remarkably linear for all three classes. A similar correlation is found by simply taking the percentage of counts in each class, as colour-coded (Supplementary Table 6). Both views provide a direct, relatively model-insensitive indication that the conformational entropy of calmodulin changes in concert with the change in the entropy of binding and that this variation can, in part, be identified with the motional class of the involved side chains.

Biological and pharmacological implications

In summary, we have employed a battery of NMR methods to characterize the dynamic response of calmodulin to the binding of six target regulatory domains. This view has been interpreted in terms of the changes in conformational entropy of calmodulin on binding. The behaviour of the six physiologically relevant interactions indicates that the conformational entropy of structured proteins can enter very significantly into high-affinity interactions between proteins. Therefore the commonly held view that high-affinity interactions are necessarily energetically dominated by specific structural (enthalpic) interactions must be relaxed to include the structural dynamics and heterogeneity that contributes to conformational protein entropy. It also seems evident that protein entropy can be exploited in the maturation of high-affinity interactions either by biological evolution or by human intervention such as in the design of protein-targeted pharmaceuticals. It will therefore be of great interest to determine how prevalent the participation of conformational entropy is across the universe of protein–protein interactions in biology. Finally, the results presented here suggest that conformational entropy can indeed play a significant part in more complex protein functions such as allostery[6].

METHODS SUMMARY

Sample preparation and isothermal titration calorimetry. Calmodulin and synthetic peptides and complexes were prepared as described previously[31] in 20 mM imidazole (pH 6.5), 100 mM KCl, 6 mM $CaCl_2$ and 0.02% (w/v) NaN_3. NMR samples were slightly (~10%) over-titrated with peptide to ensure full complex formation. For isothermal titration calorimetry, calcium-saturated calmodulin (200 μM) was used to titrate dilute solutions of peptide (5–20 μM) to avoid artefacts arising from peptide aggregation. Data were obtained with a VP-isothermal titration calorimeter (Microcal) and analysed with the Origin (v.5) software.

NMR spectroscopy. O^2_{axis} parameters were determined from T_1 and $T_{1\rho}$ deuterium relaxation[23] measured at two magnetic fields. Rotational correlation times and O^2_{NH} were determined from ^{15}N relaxation[21] obtained at two magnetic fields. $O^2_{C\alpha CO}$ parameters were determined by transverse cross-correlated relaxation[22]. All measurements were made at 35 °C. Model-free parameters[32] were determined using a grid search approach[33] using a quadrupolar coupling constant of 167 kHz, an effective N-H bond length of 1.04 Å and ^{15}N tensor breadth of 170 p.p.m. The average error of O^2_{axis}, O^2_{NH} and $O^2_{C\alpha CO}$ parameters across all complexes were estimated by Monte Carlo sampling to be 0.016, 0.011 and 0.024, respectively. The model-free parameters have been deposited in the Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu/).

Data interpretation. The change in conformational entropy of calmodulin on binding a target domain was estimated as the sum of three terms: $\Delta S_{conf} = \Delta S_{harm} + \Delta S_{rotamer(fast)} + \Delta S_{rotamer(slow)}$. ΔS_{harm} was obtained from O^2_{axis} parameters using a harmonic oscillator model[25]. Free calcium-saturated calmodulin was used as the reference state in site-to-site comparisons. To normalize the unequal number of resolved sites among the complexes, the average methyl order parameter within a complex was assigned to each unresolved site of that complex. A classical entropy term $\Delta S_{rotamer(fast)}$ was added to represent minor conformers that are sampled owing to fast rotameric interconversion[26]. For the small number of sites having multiple conformations in slow exchange on the NMR chemical shift timescale, an additional classical entropy contribution $\Delta S_{rotamer(slow)}$ was calculated on the basis of measured intensities. Populations of the three motional classes were obtained using nonlinear regression of a three gaussian model to the observed order parameter distributions.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 December 2006; accepted 25 May 2007.

1. Wodak, S. J. & Janin, J. Structural basis of macromolecular recognition. *Adv. Prot. Chem.* **61,** 9–73 (2002).
2. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267,** 383–386 (1995).
3. Spolar, R. S. & Record, M. T. Jr. Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263,** 777–784 (1994).
4. Sturtevant, J. M. Heat capacity and entropy changes in processes involving proteins. *Proc. Natl Acad. Sci. USA* **74,** 2236–2240 (1977).
5. Steinberg, I. Z. & Scheraga, H. A. Entropy changes accompanying association reactions of proteins. *J. Biol. Chem.* **238,** 172–181 (1963).
6. Cooper, A. & Dryden, D. T. F. Allostery without conformational change — a plausible model. *Eur. Biophys. J. Biophys. Lett.* **11,** 103–109 (1984).
7. Karplus, M., Ichiye, T. & Pettitt, B. M. Configurational entropy of native proteins. *Biophys. J.* **52,** 1083–1085 (1987).
8. Grunberg, R., Nilges, M. & Leckner, J. Flexibility and conformational entropy in protein–protein binding. *Structure* **14,** 683–693 (2006).
9. Igumenova, T. I., Frederick, K. K. & Wand, A. J. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem. Rev.* **106,** 1672–1699 (2006).
10. Cavanagh, J. *et al. Protein NMR spectroscopy: Principles and practice* 2nd edn (Elsevier, Burlington, Massachusetts, 2006).
11. Kahl, C. R. & Means, A. R. Regulation of cell cycle progression by calcium/calmodulin-dependent pathways. *Endocr. Rev.* **24,** 719–736 (2003).
12. Yap, K. L. *et al.* Calmodulin target database. *J. Struct. Funct. Genom.* **1,** 8–14 (2000).
13. Lee, A. L., Kinnear, S. A. & Wand, A. J. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nature Struct. Biol.* **7,** 72–77 (2000).
14. Lukas, T. J. *et al.* Calmodulin binding domains: characterization of a phosphorylation and calmodulin binding site from myosin light chain kinase. *Biochemistry* **25,** 1458–1464 (1986).
15. Zhang, M. & Vogel, H. J. Characterization of the calmodulin-binding domain of rat cerebellar nitric oxide synthase. *J. Biol. Chem.* **269,** 981–985 (1994).
16. Tokumitsu, H. *et al.* Calcium/calmodulin-dependent protein kinase kinase: identification of regulatory domains. *Biochemistry* **36,** 12823–12827 (1997).
17. Goldberg, J., Nairn, A. C. & Kuriyan, J. Structural basis for the autoinhibition of calcium/calmodulin-dependent protein kinase I. *Cell* **84,** 875–887 (1996).
18. Charbonneau, H. *et al.* Evidence for domain organization within the 61-kDa calmodulin-dependent cyclic nucleotide phosphodiesterase from bovine brain. *Biochemistry* **30,** 7931–7940 (1991).
19. Wintrode, P. L. & Privalov, P. L. Energetics of target peptide recognition by calmodulin: a calorimetric study. *J. Mol. Biol.* **266,** 1050–1062 (1997).

20. Brokx, R. D. *et al.* Energetics of target peptide binding by calmodulin reveals different modes of binding. *J. Biol. Chem.* **276,** 14083–14091 (2001).
21. Farrow, N. A. *et al.* Backbone dynamics of a free and a phosphopeptide-complexed Src homology-2 domain studied by ^{15}N NMR relaxation. *Biochemistry* **33,** 5984–6003 (1994).
22. Wang, T., Cai, S. & Zuiderweg, E. R. Temperature dependence of anisotropic protein backbone dynamics. *J. Am. Chem. Soc.* **125,** 8639–8643 (2003).
23. Muhandiram, D. R. *et al.* Measurement of H-2 T-1 and T-1p relaxation-times in uniformly C-13-Labeled and fractionally H-2-labeled proteins in solution. *J. Am. Chem. Soc.* **117,** 11536–11544 (1995).
24. Akke, M., Bruschweiler, R. & Palmer, A. G. NMR order parameters and free-energy — an analytical approach and its application to cooperative Ca^{2+} binding by calbindin-D(9k). *J. Am. Chem. Soc.* **115,** 9832–9833 (1993).
25. Li, Z., Raychaudhuri, S. & Wand, A. J. Insights into the local residual entropy of proteins provided by NMR relaxation. *Prot. Sci.* **5,** 2647–2650 (1996).
26. Lee, A. L. *et al.* Temperature dependence of the internal dynamics of a calmodulin-peptide complex. *Biochemistry* **41,** 13814–13825 (2002).
27. Lee, A. L. & Wand, A. J. Microscopic origins of entropy, heat capacity and the glass transition in proteins. *Nature* **411,** 501–504 (2001).
28. Best, R. B., Clarke, J. & Karplus, M. The origin of protein sidechain order parameter distributions. *J. Am. Chem. Soc.* **126,** 7734–7735 (2004).
29. Chou, J. J., Case, D. A. & Bax, A. Insights into the mobility of methyl-bearing side chains in proteins from $^{3}J_{CC}$ and $^{3}J_{CN}$ couplings. *J. Am. Chem. Soc.* **125,** 8959–8966 (2003).
30. Best, R. B., Clarke, J. & Karplus, M. What contributions to protein side-chain dynamics are probed by NMR experiments? A molecular dynamics simulation analysis. *J. Mol. Biol.* **349,** 185–203 (2005).
31. Kranz, J. K. *et al.* A direct test of the reductionist approach to structural studies of calmodulin activity: relevance of peptide models of target proteins. *J. Biol. Chem.* **277,** 16351–16354 (2002).
32. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **104,** 4546–4559 (1982).
33. Dellwo, M. J. & Wand, A. J. Model-independent and model-dependent analysis of the global and internal dynamics of cyclosporine-A. *J. Am. Chem. Soc.* **111,** 4571–4578 (1989).
34. Scott, D. On optimal and data-based histograms. *Biometrika* **10,** 605–610 (1979).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by a grant from the National Institutes of Health. We are grateful to S. W. Englander for helpful discussion and to Mark I. Greene for access to isothermal titration calorimetry instrumentation.

Author Contributions A.J.W. devised and initiated the project. K.K.F., M.S.M., and K.G.V. prepared the materials, collected and analysed the primary data. K.K.F. and A.J.W. performed the entropy analysis. A.J.W. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.J.W. (wand@mail.med.upenn.edu).

METHODS

Data interpretation. The summed distribution of O^2_{axis} parameters of all six complexes was fitted to a random distribution and to one, two- and three-gaussian distribution models. The bin size for this analysis was determined from a well-established formula for optimal bin width[34] and was found to be 0.05. Only the three-gaussian model could satisfactorily describe the data ($P < 0.0001$), that is:

$$\text{Occurrences}\,(O^2_{axis}) = A_J \exp[((-O^2_{axis} - C_J)/W_J)^2/2] + A_\alpha \exp[((-O^2_{axis} - C_\alpha)/W_\alpha)^2/2] + A_\omega \exp[((-O^2_{axis} - C_\omega)/W_\omega)^2/2] \quad (3)$$

where A_i, C_i and W_i define the population, centre and breadth, respectively, of the J, α and ω classes of motion.

The nine parameters were fitted using the nonlinear regression routine of SigmaPlot 2000 (SPSS). The summed distribution yielded fitted 3-gaussian distributions centred on O^2_{axis} values of 0.35 (J-class), 0.58 (α-class) and 0.78 (ω-class). These centres were fixed in subsequent fitting of the O^2_{axis} distributions of the individual complexes, from which the relative populations of each motional class were obtained. Uncertainties in the fitted populations were estimated by varying the O^2_{axis} parameters by two standard deviations. This results in asymmetric error bars. The total change in conformational entropy of calmodulin on binding a target domain was calculated as the sum of three terms: $\Delta S_{conf} = \Delta S_{harm} + \Delta S_{rotamer(fast)} + \Delta S_{rotamer(slow)}$. Changes in conformational entropy $\Delta S_{rotamer(fast)}$ expressed as changes in motion within a rotameric well on the fast timescale (sub-ns) were obtained from the experimentally determined O^2_{axis} parameters using a simple harmonic oscillator model[25]. To calculate changes in entropy derived from motion of the same oscillator, site-to-site comparison to free calcium-saturated calmodulin was used to provide a reference state. The change in the entropy reflected by the change in the motion of each methyl symmetry axis was estimated using $\Delta S_{harm} = -18 \times \Delta O^2_{axis}$ J mol^{-1} K^{-1}. See ref. 25 for further details of the model. To normalize the unequal number of resolved sites among the complexes, the average methyl order parameter within a complex was assigned to each unresolved site of that complex. A classical entropy term ($\Delta S_{rotamer(fast)}$) was added to represent minor conformers that are sampled owing to fast rotameric interconversion that also contributes to the generalized order parameter[26]. For the small number of methyl sites having multiple conformations in slow exchange on the NMR chemical shift timescale, an additional classical entropy contribution ($\Delta S_{rotamer(slow)}$) was calculated using the measured intensities to provide populations. See Supplementary Information for additional details and results.

LETTERS

The sources of sodium escaping from Io revealed by spectral high definition imaging

Michael Mendillo[1,2], Sophie Laurent[1,2], Jody Wilson[1], Jeffrey Baumgardner[1], Janusz Konrad[2] & W. Clem Karl[2]

On Jupiter's moon Io, volcanic plumes and evaporating lava flows provide hot gases to form an atmosphere that is subsequently ionized. Some of Io's plasma is captured by the planet's strong magnetic field to form a co-rotating torus at Io's distance; the remaining ions and electrons form Io's ionosphere. The torus and ionosphere are also depleted by three time-variable processes that produce a banana-shaped cloud orbiting with Io[1], a giant nebula extending out to about 500 Jupiter radii[2–5], and a jet close to Io[6–9]. No spatial constraints exist for the sources of the first two; they have been inferred only from modelling the patterns seen in the trace gas sodium observed far from Io. Here we report observations that reveal a spatially confined stream that ejects sodium only from the wake of the Io–torus interaction, together with a visually distinct, spherically symmetrical outflow region arising from atmospheric sputtering. The spatial extent of the ionospheric wake that feeds the stream is more than twice that observed by the Galileo spacecraft and modelled successfully. This implies considerable variability, and therefore the need for additional modelling of volcanically-driven, episodic states of the great jovian nebula.

Although neutral atoms of sodium are a minor constituent in Io's volcanic emissions, they have bright spectral lines (589.0 and 589.6 nm) in the visible portion of the spectrum accessible to Earth-based telescopes, and thus they serve as the key tracer of the overall neutral-plasma budget in the Jupiter–Io system. Tempering this view is the fact that the total amount of sodium is low and thus very faint. To observe Na coronal and stream distributions on a spatial scale of a few Io radii (R_{Io}), one must conduct low-light-level imaging in a narrow spectral band close to a bright target (Io) that is itself close to an even brighter object (Jupiter). The observational challenges involve minimizing scattered light into a telescope and within its optics, coping with the random image motions of the target due to turbulence in the Earth's atmosphere (seeing), and the low photon count from the narrow wavelength band needed to isolate the sodium emission lines. The Boston University high-definition-imaging (HDI) system was developed specifically for this task of obtaining high quality spectral images[10] of a low-light-level target.

The HDI approach was designed for use on telescopes without adaptive optics. It takes advantage of the statistical nature of turbulence: every now and then a remarkably clear image flashes into sight among a flood of jittering, fuzzy images. Past studies[11] have examined the probability of obtaining such a 'lucky image' and described ways to optimize success. Turbulence involves variations in atmospheric density that bend light waves, thereby moving and distorting the image. If the size of the density cells is larger than the aperture of the telescope, the diffraction is fairly uniform across the lens or mirror and the seeing for that instant is good; if the aperture is much larger than the turbulence cells, then many distortions of the image occur across it at the same time and the seeing is poor. Thus, diffraction-limited images are more easily obtained with short exposures using small telescopes.

For a large and bright target (for example, Mercury) observed in white light (that is, over a broad spectral band), a single 'best image' can be found from a series of rapid time exposures[12]. An improvement is possible if many good images distributed in a data set can be found and added together. HDI creates such a composite image by selecting and adding temporally spaced images to yield a high signal-to-noise ratio (SNR) under diffraction-limited conditions. This is done in a post-processing stage using data sets taken in burst-mode (high-rate image acquisition). The three-step automated method is: (1) find the centre of each image in a long data stream, (2) down-select to the very few high-quality images, and then (3) shift-and-add them. When done for Mercury in white light, the finest image of its surface ever obtained from a ground-based telescope was obtained by co-adding the 60 'best images' from a total of 190,000 images[10,13].

Attempting the above method in spectral mode (for example, using a filter to select a narrow wavelength band) fails because the short exposures required to 'freeze the seeing' do not allow enough photons into the detector to obtain an image with an SNR sufficient for post-processing. To solve this problem, our HDI instrument passes light from a celestial object through a beam splitter, with 50% recorded as a white-light image (high SNR), and 50% passed to an image slicer[14] for the creation of images in the spectral domain (each with a low SNR). In both cases, the data are recorded simultaneously at video rates. All of the registration and selection is done using the white-light images and then this information is passed into the spectral domain for shifting and adding of only the selected images, those that were themselves too faint to support independent image processing.

To get sufficient SNR and to have spatial resolution to resolve features on the scale of Io radii, a telescope with a large aperture was needed. This counteracts the advantages described for the high probability of good images from small telescopes[11]. Our trade-off was to select a large telescope at a site known for its atmospheric transparency, to use short exposures (one-sixtieth of a second) to freeze the seeing, and to correct for the remaining (inevitable) turbulence-induced problems by use of our HDI post-processing techniques.

We used our HDI instrument (see Methods) on the Advanced Electro-Optical System (AEOS) 3.67 m telescope operated by the US Air Force on Mt Haleakala, Hawaii, in December 2000. Our image acquisition scheme takes and stores 7 min of observations, collecting 12,500 images in both white light and in spectral light simultaneously. Five such data bursts were taken over a 90 min span, resulting in 62,500 images. The HDI protocols of registration, shifting and adding that were previously developed for a large target in white light (Mercury)[10] required new methods for a small faint target (Io). The technique developed[14,15] involved scanning a template (an electronic mask the angular size of Io at the time of observation) over each

[1]Center for Space Physics, [2]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215, USA.

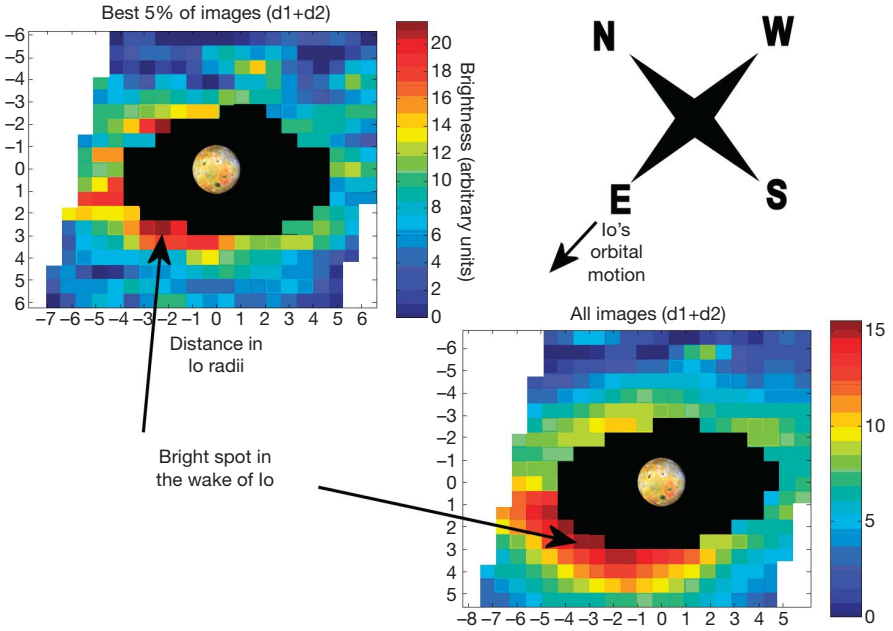


Figure 1 | The distribution of sodium near Io reveals distinctive sources. Observations taken on 4 December 2000 at the AEOS telescope in Maui, Hawaii, between 13:04 and 14:35 UT, are shown on a spatial scale of radii of Io. The colour-coded brightness levels are in arbitrary units. The irregularly shaped dark region surrounding Io denotes areas where scattered light within the spectrograph precluded the complete subtraction of on-band versus off-band light[14]. A composite image from the 5% best-seeing images identifies a localized source (lower left) and a symmetrical source. The second image results from the registration, shifting and adding of all 62,000 images using the Na (d1+d2) emission lines. The orientation key gives the sky-view from Earth, showing that the brightest pixels are in the eastward direction, down-stream of torus flow past Io in its orbit. See Fig. 2 for additional geometrical information.

digital array containing Io (fuzzy in white light) to locate their centres (that is, by finding the position that minimized total brightness beyond the template). After all the images were registered to a common centre, the integrated brightness levels in a belt of constant radial width beyond the template were ordered from low (image with least scattered light, that is, best seeing) to high (worst seeing). These provided the selection criteria for adding images of like quality. In Fig. 1 we show two HDI product images, one using the top 5% best-seeing frames (~3,000 frames × 1/60 s = 50 s of total exposure time). The second comes from centring and adding all 62,500 frames, thus having 35 min of exposure time.

The distribution of Na about Io in Fig. 1 clearly has two components. There is a symmetrical region of faint emission surrounding the moon. The second and more prominent feature is the bright and non-symmetrical emission in the lower-left portion of each image. To understand the origins of these spatially distinct sources, Fig. 2 gives a schematic portrayal of the three-dimensional nature of the Io–plasma torus interaction. Jupiter's plasma torus co-rotates with the planet at the same distance and in the same direction as Io's orbit, but at a speed of 75 km s^{-1} in comparison to Io's orbital speed of 17 km s^{-1}. Thus, a high-speed plasma wind continuously overtakes the moon and has two types of interactions with Io's atmosphere: (1) torus ions collide with neutral Na atoms to create an always present 'sputtering source' of ejected Na, and (2) the magnetized plasma wind slows down and captures and/or distorts most of the tiny moon's ionosphere as it sweeps by, except in the region physically shielded from the wind. That is, Io's extended and stagnant ionosphere exists primarily in the down-stream or Io-wake side of the torus[16–21].

The first effect (atmospheric sputtering) accounts for Io's innermost neutral Na pattern, the so-called banana cloud (far larger than the scale of the image in Fig. 1), as well as for the most quiescent state of the great sodium nebula when Io's volcanoes have minimal activity[4,5]. We thus associate the symmetrical Na coronal distribution in Fig. 1 with the atmospheric sputtering source. Episodic brightenings of the nebula occur when volcanic-induced molecular ions (for example, $NaCl^+$) are added to Io's ionosphere[5,22–25]. The second effect (torus capture of ionospheric plasma) leads to dissociative recombinations of these ions with electrons ($NaCl^+ + e^- \rightarrow Na + Cl$) and thus to an extended Na stream source. We associate the spatially distinct Na distribution in Fig. 1 with the ionospheric base of the stream source.

There are two aspects of the stream source—its temporal variability and its Io wake location—that need addressing to validate our conclusion. Although no volcanic activity was monitored during the period of our observations, we did make observations of the large-scale sodium nebula one month before and after this AEOS data set. We found a nebula caused by a moderate stream source in November 2000, changing to a weaker stream and sputter-dominated source in

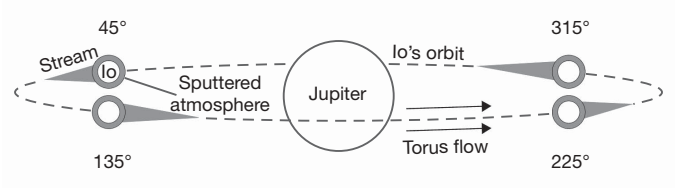


Figure 2 | A spatial relationship exists between a terrestrial view of target locations at Jupiter. This schematic (not to scale) shows the Earth-based view of the orbital configuration of Jupiter, Io and the two distinct emission regions presented in Fig. 1. A symmetrical region from atmospheric sputtering is indicated for all orbital positions. An additional stream source can occur in the wake of the torus flow. If there is a Jupiter-facing hemispheric enhancement of Io's sodium corona, a hemispheric asymmetry would appear on the same side as the stream source at orbital phase angles of ~135° and 315°, but in opposite directions at phase angles ~45° and 225°. The observations in Fig. 1 at ~315° show that the stream source dominates any coronal asymmetry by virtue of its narrow emission pattern away from Io. It is also distinct from a potential jet source[6] that typically points away from Jupiter at all phase angles and has a brightness pattern that extends a few jovian radii from Io, in marked contrast to the emission signature of the stream source in Fig. 1.

January 2001 (ref. 26). The image in Fig. 1 suggests that the stream source was still operating at a low but detectable rate during the intervening month of December 2000.

A second issue arises from the fact that volcanic activity is dominant on Io's Jupiter-facing hemisphere, and that Io's Na corona was observed to have enhanced sodium emission on that side[27]. As shown in Fig. 2, Io was at about half its maximum angular separation (elongation angle) from Jupiter at the time of our observations, in the portion of its orbit about to pass behind the planet. At this orbital phase angle, it is impossible visually to distinguish between a Jupiter-facing coronal asymmetry and a stream feature of comparable extent. However, the HDI results using the 5% best seeing images isolate a narrow region (1 to 2 pixels) of peak emission, and one that remains far from hemispheric when all images are used. In addition, subsequent observations were conducted at the phase angle associated with Io's emergence from behind Jupiter, when a Jupiter-facing coronal feature and a torus-wake feature would be in opposite directions. Finding only the wake feature to the east confirmed the reality and persistence of the stream's characteristic signature in the discovery image shown in Fig. 1.

The results presented here show that the most variable source of sodium escaping from Io coincides unambiguously with the location of Io's distorted and nearly stagnant wake-side ionosphere. Figure 1 sets specific dimensions to this feature. Sodium comes from where the ionosphere is dense, and thus its dilution to high-speed co-rotational flow occurs where the base of the Na stream matches the sputtering-induced corona (indicated by the light-blue colour coding). The extent of this stagnation distance, when projected to Io's orbit, comes to about six Io radii, as reported by radio occultation methods[21]. Prior modelling[19] of the dimensions of the wake based on *in situ* Galileo data[20] depict the ionospheric stagnation region to be less than three Io radii in extent. Figure 1 thus offers new goals for models treating the two-dimensional spatial and temporal variability of processes that add plasmas and neutrals to the jovian magnetosphere.

METHODS

The HDI hardware is an 'image slicer', a three-dimensional spectrograph (giving brightness versus two spatial dimensions and wavelength). A target is imaged upon a 20×20 pixel array coupled to 400 optical fibres at the image plane end that are re-configured at the other end into a linear bundle of 400 elements. This linear representation of a two-dimensional image then feeds a traditional spectrograph (giving brightness in space versus wavelength). This results in the full spectral signatures of a modified portrayal of space. Mapping the 400 elements at each wavelength back to two-dimensional space results in a 'data cube' of images at multiple wavelengths. Selection of a 'slice' of this cube gives a spectral image from $\sim$1.5 Å centred on the d1 and d2 sodium lines.

Received 9 April; accepted 1 June 2007.

1. Smyth, W. H. & Combi, M. R. A general model for Io's neutral gas clouds. II. Application to the sodium cloud. *Astrophys. J.* **328,** 888–918 (1988).
2. Schneider, N. M. *et al.* Molecular origin of Io's fast sodium. *Science* **253,** 1394–1397 (1991).
3. Mendillo, M., Baumgardner, J., Flynn, B. & Hughes, J. The extended sodium nebula of Jupiter. *Nature* **348,** 312–314 (1990).
4. Wilson, J. K. *et al.* The dual sources of Io's sodium clouds. *Icarus* **157,** 476–489 (2002).
5. Mendillo, M., Wilson, J., Spencer, J. & Stansbury, J. Io's volcanic control of Jupiter's extended neutral clouds. *Icarus* **170,** 430–442 (2004).
6. Pilcher, C. B., Fertel, J. H., Smyth, W. H. & Combi, M. R. Io's sodium directional features—Evidence for a magnetospheric-wind-driven gas escape mechanism. *Astrophys. J.* **287,** 427–444 (1984).
7. Goldberg, B. A., Garneau, G. W. & Lavoie, S. K. Io's sodium cloud. *Science* **226,** 512–516 (1984).
8. Wilson, J. K. & Schneider, N. M. Io's sodium directional feature: Evidence for ionospheric escape. *J. Geophys. Res.* **104,** 16567–16584 (1999).
9. Burger, M. H. Schneider, N. M. & Wilson, J. K. Galileo's close-up view of the Io sodium jet. *Geophys. Res. Lett.* **26,** 3333–3336 (1999).
10. Baumgardner, J., Mendillo, M. & Wilson, J. K. A digital high definition imaging system for spectral studies of extended planetary atmospheres. 1. Initial results in white light showing features on the hemisphere of Mercury unimaged by Mariner 10. *Astron. J.* **119,** 2458–2464 (2000).
11. Fried, D. L. Probability of getting a lucky short-exposure image through turbulence. *J. Opt. Soc. Am.* **68,** 1651–1658 (1978).
12. Warrel, J&. Limaye, S. S. Properties of the hermean regolith: I. Global regolith albedo variation at 200 km scale from multicolor CCD imaging. *Planet. Space Sci.* **49,** 1531–1552 (2001).
13. Mendillo, M. *et al.* Imaging the surface of Mercury using ground-based telescopes. *Planet. Space Sci.* **49,** 1501–1505 (2001).
14. Laurent, S. *Design of a High Definition Imaging (HDI) Analysis Technique Adapted to Challenging Environments* PhD dissertation (College of Engineering, Boston University, Boston, 2004); ⟨http://sirius.bu.edu/aeronomy/laurentthesis.pdf⟩.
15. Laurent, S. *et al.* Design of a High Definition Imaging (HDI) analysis technique adapted to challenging environments. In *Applications of Digital Image Processing XXVII* (ed. Tescher, A. G.) *Proc. SPIE* **5558,** 676–687 (2004).
16. Linker, J. A., Kivelson, M. G. & Walker, R. J. A three-dimensional MHD simulation of plasma flow past Io. *J. Geophys. Res.* **96,** 21037–21053 (1991).
17. Saur, J., Neubauer, F. M., Strobel, D. F. & Summers, M. E. Three-dimensional plasma simulation of Io's interaction with the Io plasma torus: Asymmetric plasma flow. *J. Geophys. Res.* **104,** 25105–25126 (1999).
18. Combi, M. R., Kabin, K., Gombosi, T. I., DeZeeuw, D. L. & Powell, K. G. Io's plasma environment during the Galileo flyby: Global three-dimensional MHD modeling with adaptive mesh refinement. *J. Geophys. Res.* **103,** 9071–9082 (1998).
19. Kabin, K. *et al.* Io's magnetospheric interaction: an MHD model with day-night asymmetry. *Planet. Space Sci.* **49,** 337–344 (2001).
20. Frank, L. A. *et al.* Plasma observations at Io with the Galileo spacecraft. *Science* **274,** 394–395 (1996).
21. Hinson, D. P. *et al.* Galileo radio occultation measurements of Io's ionosphere and plasma wake. *J. Geophys. Res.* **103,** 29343–29357 (1998).
22. Kuppers, M. & Schneider, N. M. Discovery of chlorine in the Io torus. *Geophys. Res. Lett.* **27,** 513–516 (2000).
23. Lellouch, E., Paubert, G., Moses, J. J., Schneider, N. M. & Strobel, D. F. Volcanically emitted sodium chloride as a source for Io's neutral clouds and plasma torus. *Nature* **42,** 45–47 (2003).
24. Zolotov, M. Yu & Fegley, B. Jr. Eruption conditions of Pele volcano on Io inferred from chemistry of its volcanic plume. *Geophys. Res. Lett.* **27,** 2789–2792 (2000).
25. Moses, J. J., Zolotov, M. Yu & Fegley, B. Jr. Alkali and chlorine photochemistry in a volcanically driven atmosphere on Io. *Icarus* **156,** 107–135 (2002).
26. Wilson, J. K. & Mendillo, M. A change in Io's sodium clouds during the Cassini/Jupiter flyby. *Eos* **82,** S255 (2001).
27. Burger, M. H. *et al.* Mutual event observations of Io's sodium corona. *Astrophys. J.* **563,** 1063–1074 (2001).

Acknowledgements We thank P. Kervin and staff of the AEOS installation for their expertise and collegiality in the use of our HDI system for these measurements. The Boston University HDI image slicer was built under a Defense University Research Instrumentation Program (DURIP) grant via sponsorship by the Office of Naval Research (ONR). Observations were made possible by a grant from the joint NSF/AFOSR programme for AEOS, and data analysis was funded by a grant from the NSF Planetary Astronomy Program.

Author Contributions The observations were made by J.B., J.W. and M.M., who also guided the overall study. S.L. developed the HDI techniques and conducted the data analysis with methodological expertise provided by J.K, W.C.K., J.B. and J.W. M.M. wrote most of the paper, with contributions and interpretation of the data from J.W. and J.B.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M. (mendillo@bu.edu).

LETTERS

Interference between two indistinguishable electrons from independent sources

I. Neder[1], N. Ofek[1], Y. Chung[2], M. Heiblum[1], D. Mahalu[1] & V. Umansky[1]

Very much like the ubiquitous quantum interference of a single particle with itself[1], quantum interference of two independent, but indistinguishable, particles is also possible. For a single particle, the interference is between the amplitudes of the particle's wavefunctions, whereas the interference between two particles is a direct result of quantum exchange statistics. Such interference is observed only in the joint probability of finding the particles in two separated detectors, after they were injected from two spatially separated and independent sources. Experimental realizations of two-particle interferometers have been proposed[2,3]; in these proposals it was shown that such correlations are a direct signature of quantum entanglement[4] between the spatial degrees of freedom of the two particles ('orbital entanglement'), even though they do not interact with each other. In optics, experiments using indistinguishable pairs of photons encountered difficulties in generating pairs of independent photons and synchronizing their arrival times; thus they have concentrated on detecting bunching of photons (bosons) by coincidence measurements[5,6]. Similar experiments with electrons are rather scarce. Cross-correlation measurements between partitioned currents, emanating from one source[7–10], yielded similar information to that obtained from auto-correlation (shot noise) measurements[11,12]. The proposal of ref. 3 is an electronic analogue to the historical Hanbury Brown and Twiss experiment with classical light[13,14]. It is based on the electronic Mach–Zehnder interferometer[15] that uses edge channels in the quantum Hall effect regime[16]. Here we implement such an interferometer. We partitioned two independent and mutually incoherent electron beams into two trajectories, so that the combined four trajectories enclosed an Aharonov–Bohm flux. Although individual currents and their fluctuations (shot noise measured by auto-correlation) were found to be independent of the Aharonov–Bohm flux, the cross-correlation between current fluctuations at two opposite points across the device exhibited strong Aharonov–Bohm oscillations, suggesting orbital entanglement between the two electron beams.

In many ways, experiments with electrons are easier than those with photons. Injecting electrons from an extremely cold and degenerate fermionic reservoir produces a highly ordered beam of electrons that is totally noiseless[17]; hence, a high coincidence rate is achieved without the need to synchronize the arrival times of the electrons. As each electron has a definite energy (Fermi energy) and momentum (Fermi momentum), electrons can be made indistinguishable by injecting them from two equal voltage sources. Moreover, because the coherence length of the electrons ('wave packet width' or 'spatial size') is determined by the source voltage (at low temperature), a very small source voltage ensures the presence of a single electron at a time in the interferometer, preventing electron–electron interactions. However, the small voltage leads to an exceedingly small electrical current and to minute fluctuations, making the measurements extremely difficult to perform.

A diagram of our experiment is shown in Fig. 1a (ref. 2). Two independent, separated, sources of electrons (S1 and S2) inject ordered, hence noiseless, electrons towards each other. Each stream passes through a beam splitter (A and B), and splits into two negatively correlated partitioned streams (if an electron turns right, a hole is injected to the left). Both sets of the two partitioned streams join each other at two additional beam splitters (C and D), interfere there and generate altogether four streams that are collected by drains D1–D4. Hence, every electron emitted by either S1 or S2 eventually arrived at one of the four drains. Consider now the event where one electron arrives at D2 and the other arrives at D4. There are two quantum mechanical probability amplitudes contributing to this event: S1 to D2 and S2 to D4; or, alternatively, S1 to D4 and S2 to D2. These two 'two-particle' events can interfere because they are indistinguishable. Because in the two possible events the electrons travel along different paths (thus accumulating different phases), the joint probability of one arriving at D2 and the other at D4 contains the total phase of all paths—as we show below.

The two wavefunctions, corresponding to the incoming states from each of the two sources Ψ_{Si}, can be expressed in the basis of the outgoing states at the four drains ψ_{Dj}. Assuming, as in the experiment, that every beam splitter is half reflecting and half transmitting, its unitary scattering matrix M (that ties the input and output states) can be taken as: $M=\begin{bmatrix} r & t \\ t' & r' \end{bmatrix}=\frac{1}{\sqrt{2}}\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}$. Considering the phases of the four possible paths $\phi_1,..,\phi_4$:

$$\Psi_{S1}(x)=\frac{1}{2}\left[ie^{i\phi_1}\psi_{D1}(x)-e^{i\phi_1}\psi_{D2}(x)+ie^{i\phi_2}\psi_{D3}(x)+e^{i\phi_2}\psi_{D4}(x)\right] \quad (1a)$$

$$\Psi_{S2}(x)=\frac{1}{2}\left[ie^{i\phi_3}\psi_{D1}(x)+e^{i\phi_3}\psi_{D2}(x)+ie^{i\phi_4}\psi_{D3}(x)-e^{i\phi_4}\psi_{D4}(x)\right] \quad (1b)$$

As, in this set-up, each electron is not allowed to interfere with itself, only particle statistics could cause interference. Because of the fermionic property of electrons, the total two-particle wavefunction must be the antisymmetric product of equation (1a) and equation (1b):

$$\Psi_{total}(x_1,x_2)=\frac{1}{\sqrt{2}}\left[\Psi_{S1}(x_1)\,\Psi_{S2}(x_2)-\Psi_{S2}(x_1)\,\Psi_{S1}(x_2)\right] \quad (2)$$

with x1 and x2 any two locations in the interferometer. Substituting equation (1) in equation (2) leads to 24 terms, expressing the probability amplitude for one electron at x1 and another at x2. As we wish to concentrate on correlations between drains, we write Ψtotal using the notation $\psi_{DiDj}\equiv\frac{1}{\sqrt{2}}\left[\psi_{Di}(x_1)\,\psi_{Dj}(x_2)-\psi_{Dj}(x_1)\,\psi_{Di}(x_2)\right]$ for an antisymmetric state, in which one electron heads to Di and another

[1]Braun Center for Submicron Research, Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Department of Physics, Pusan National University, Busan 609-735, Korea.

to Dj. The two-particle wavefunction is:

$$\Psi_{\text{total}}(x_1,x_2)=\frac{i}{2}\left(e^{i(\phi_1+\phi_3)}\psi_{\text{D1D2}}-e^{i(\phi_2+\phi_4)}\psi_{\text{D3D4}}\right)+$$
$$+\frac{i}{2}e^{\frac{i}{2}(\phi_1+\phi_2+\phi_3+\phi_4)}\left[\sin\left(\frac{\Phi_{\text{total}}}{2}\right)(\psi_{\text{D2D4}}-\psi_{\text{D1D3}})-\cos\left(\frac{\Phi_{\text{total}}}{2}\right)(\psi_{\text{D2D3}}+\psi_{\text{D1D4}})\right] \quad (3)$$

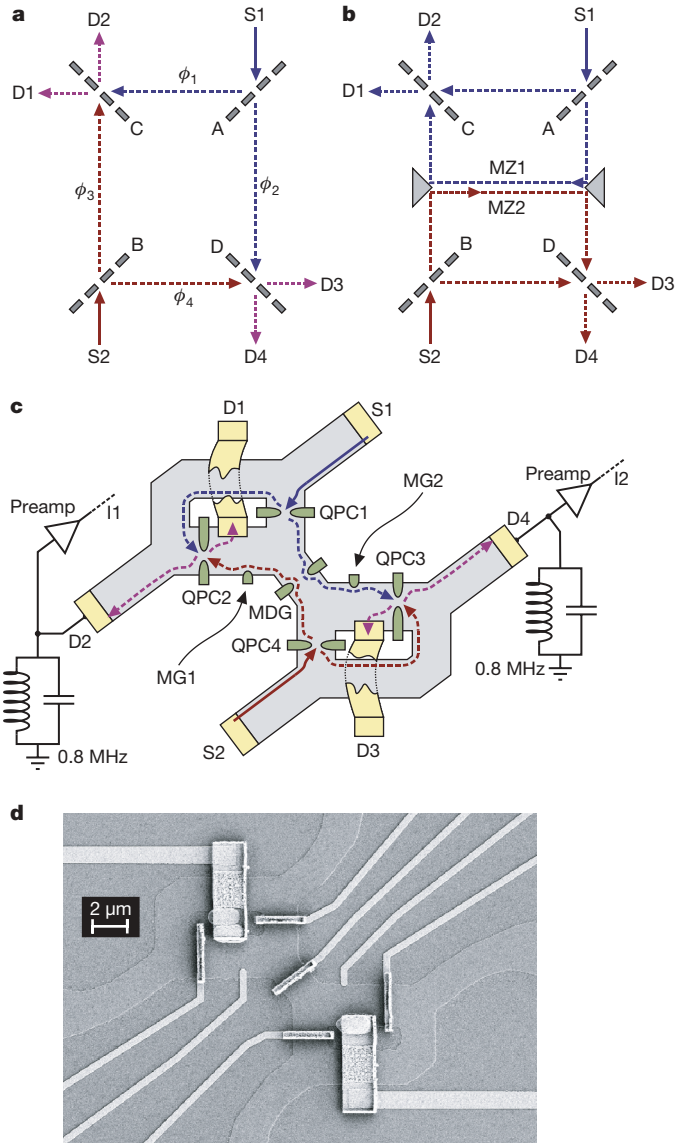


Figure 1 | The two-particle Aharonov–Bohm interferometer. a, Diagram of the interferometer. Sources S1 and S2 inject streams of particles, which are split by beam splitters A and B, later to recombine at beam splitters C and D. Each particle can arrive at any of four different drains, D1–D4. Each of the four trajectories accumulates phase ϕ_i. **b**, By breaking the interferometer in the centre, two separate Mach–Zehnder interferometers (MZIs) are formed. The MZIs are the building blocks of the two-particle interferometer. **c**, A detailed drawing of the interferometer. It was fabricated on a high mobility GaAs-AlGaAs heterostructure, with a two-dimensional electron gas buried some 70 nm below the surface (carrier density 2.2×10^{11} cm^{-2} and low temperature mobility 5×10^{6} cm^2 V^{-1} s^{-1}). Samples were cooled to ~10 mK electron temperature. Quantum point contacts (QPCs) served as beam splitters, and ohmic contacts as sources and drains. Tuning gates MG1 and MG2 changed the area and thus the magnetic flux threaded through the interferometer (at filling factor one of the integer quantum Hall effect). 'Middle gate' MDG separated the interferometer into two MZIs. Metallic air bridges connected drains D1 and D3 to the outside, where they were grounded. Currents at D2 and D4 were filtered first by an *LC* circuit (tuned to 0.8 MHz and 60 kHz bandwidth) and then amplified by a cold preamplifier (at 4.2 K). **d**, Scanning electron micrograph of the actual sample. Air bridges were used to contact the small ohmic contacts, the split gates of the QPCs, and the MDG.

with $\Phi_{\text{total}}=\phi_1-\phi_3+\phi_4-\phi_2$, which is exactly the total accumulated phase going anti-clockwise along the four trajectories of the two particles.

Equation (3) describes the two-particle interference effect, with the absolute value squared of the prefactor of $\psi_{\text{D}i\text{D}j}$ the joint probability of having one electron at D*i* and one at D*j*. Concentrating on the correlation between D2 and D4, one can deduce from equation (3) the following: (1) two electrons never arrive at the same drain (Pauli exclusion principle); (2) the first part suggests that there is a 50% chance for two electrons to arrive at the same 'side' simultaneously, namely, at D1 and D2, or at D3 and D4, but never at D2 and D4; (3) the second part suggests that there is a 50% chance for two electrons to arrive at opposite 'sides', namely, one at D1 or D2 and the other at D3 or D4; however, the exact correlation depends on Φ_{total}. When $\Phi_{\text{total}}=\pi$, $\sin^2(\Phi_{\text{total}}/2)=1$ and two electrons arrive at (D1, D3) or at (D2, D4), but when $\Phi_{\text{total}}=0$, $\cos^2(\Phi_{\text{total}}/2)=1$ and the complementary events take place. (4) Combining all events in the two parts of the total wavefunction, one finds for $\Phi_{\text{total}}=0$ a perfect anti-correlation between the arrival of electrons in D2 and in D4; however, for $\Phi_{\text{total}}=\pi$ there is 50% chance of anti-correlation (first part) and 50% chance of positive correlation (second part)—hence, zero correlation. The time-averaged cross-correlated signal of the current fluctuations in the two drains is proportional to the probability of the correlated arrival of electrons in these drains. Varying the total phase should result in a negative oscillating cross-correlation signal between current fluctuations in D2 and in D4. The quantitative estimate of the amplitude of that cross-correlation signal is discussed later.

Figure 1b describes the realization of the experiment. The two-particle interferometer is shown split in the centre, resulting in an upper and lower segments; each is a simple optical Mach–Zehnder interferometer (MZI)[18]. An electronic version of the MZI has been recently fabricated and studied[15,19,20]. A quantizing magnetic field (~6.4 T) brings the two-dimensional electron gas into the quantum Hall effect state at filling factor one. The current is carried by a single edge channel along the boundary of the sample[16]. Being a chiral one-dimensional object, the channel is highly immune to back scattering and dephasing. The layout of the two-particle interferometer is described in Fig. 1c, with the scanning electron micrograph of the actual device shown in Fig. 1d. The two MZIs can be separated from each other with a 'middle gate' (MDG). When it is closed, each MZI can be tested independently for its coherence and the Aharonov–Bohm periodicity. A quantum point contact (QPC), formed by metallic split gates, functions as a beam splitter while ohmic contacts serve as sources and drains. In this configuration, the phase that is accumulated along the four trajectories is the Aharonov–Bohm phase (φ_{AB}), namely, $\Phi_{\text{total}}=\varphi_{\text{AB}}=2\pi BA/\Phi_0$, with B the magnetic field and A the area enclosed by the four paths ($\Phi_0=4.14\times10^{-15}$ T m^2 is the flux quantum)[21]. Look, for example, at the upper MZI of the separated two-particle interferometer (Fig. 1b). An edge channel, emanating from ohmic contact S1, is split by QPC1 into two paths that enclose a high magnetic flux and join again at QPC2. The phase dependent transmission coefficient from S1 to D2 is:

$$T_{\text{MZI}}=\left|t_{\text{QPC1}}t_{\text{QPC2}}+e^{i\varphi_{\text{AB}}}r_{\text{QPC1}}r_{\text{QPC2}}\right|^2=T_0+T_\varphi\cos(\varphi_{\text{AB}}) \quad (4)$$

where t and r are the transmission and reflection amplitudes of the QPCs. The visibility is defined as the ratio between the phase-dependent and the phase-independent terms, $v_{\text{MZI}}=T_\varphi/T_0$. The Aharonov–Bohm phase was controlled by the magnetic field and the 'modulation gate' (MG1 or MG2) voltage V_{MG}, which affected the area enclosed by the two paths.

Figure 2 displays the measured conductance of the two separated MZIs (defined as $i_{\text{D}}/V_{\text{S}}=T_{\text{MZI}}(e^2/h)$, where i_{D} is the AC current in the drain, V_{S} the applied a.c. voltage at the sources, with e^2/h the edge channel conductance). Pinching off MDG, the QPCs were tuned to transmission 0.5 and the AC signal was measured at D2 and D4 as a function of V_{MG} and magnetic field. As V_{MG} was scanned repeatedly

the magnetic field decayed unavoidably (as the superconducting magnet is not ideal) at a rate of $\sim$1.4 G h^{-1}. Hence, the interference pattern was 'tilted' in the two-coordinate plane of V_{MG} and time (magnetic field), with two basic Aharonov–Bohm periods for each MZI[15]. Apparently, the seemingly identical MZIs had different periodicities: 1 mV and 80 min in the upper MZI, and 1.37 mV and 87 min in the lower MZI (the asymmetry resulted from misaligning the QPCs and modulation gates). In the two MZIs, we found visibilities 75–90%, by far the highest measured in an electron interferometer. The high visibility was likely to result from the smaller size of the MZIs[15,19,20]; hence, dephasing mechanisms such as flux fluctuations or temperature smearing were less effective. Moreover, the high quality two-dimensional electron gas assured a better formation of one-dimension-like edge channels and better overlap of particle wavefunctions.

We then discharged MDG, thus opening it fully and turning the two MZIs into a single two-particle interferometer. The conductances at D2 and D4 were now found to be independent of the Aharonov–Bohm flux, with a visibility smaller than the background ($<$0.1%). This is expected, as each electron did not enclose an Aharonov–Bohm flux any more.

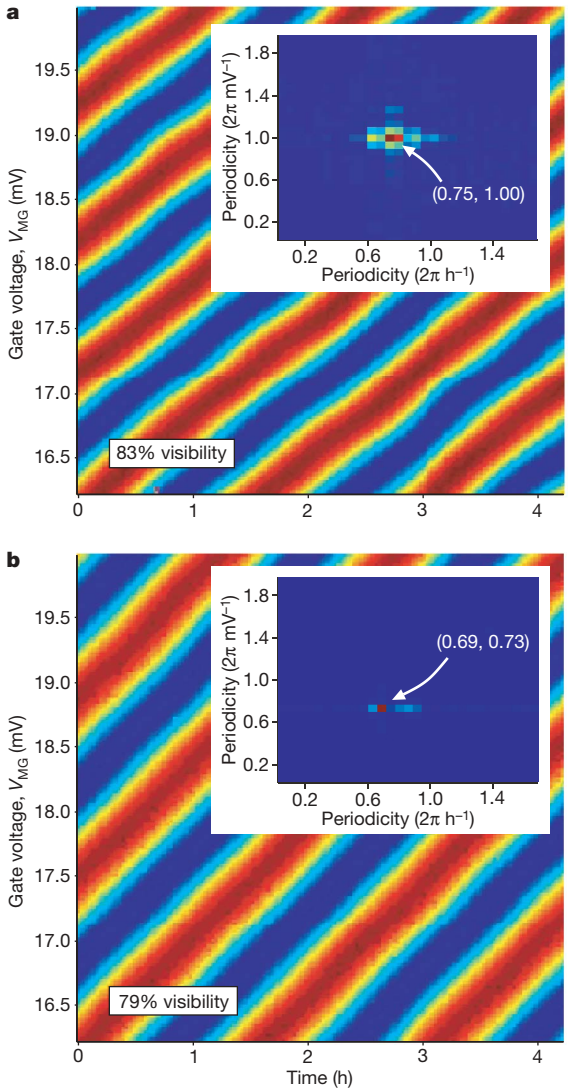


Figure 2 | Colour plot of the conductance of the two separate MZIs as function of the modulation gate voltage and the magnetic field that decayed in time. Strong Aharonov–Bohm oscillations dominate the conductance with visibilities of $\sim$80% each. A two-dimensional FFT in the inset provides the periodicity in modulation gate voltage (V_{MG}) and in time.

We turn now to discuss the current fluctuations, namely, the shot noise in D2 and in D4. Feeding a d.c. current into S1, the low frequency spectral density of the shot noise in the partitioned current (by QPC1) at D2 and at D4 (with QPC2 closed and QPC3 and MDG open) was measured. Its expected value (neglecting here finite temperature corrections) is $S_{D2} = 2eI_{S1}T_{QPC1}(1 - T_{QPC1}) = 0.5eI_{S1}$ (A^2 Hz^{-1}) for $T_{QPC1} = |t_{QPC1}|^2 = 0.5$ (ref. 15). The current fluctuations in the drain were filtered by an *LC* circuit, with 60 kHz bandwidth around a centre frequency $\sim$0.8 MHz, and then amplified by the cold amplifier, followed by a room-temperature amplifier and a spectrum analyser. In order to calibrate the cross-correlation measurement, we performed three noise measurements: (1) noise measured at D2; (2) noise measured at D4; and (3) noise measured by cross-correlating the current fluctuations at D2 and at D4 (by an analogue home-made correlating circuit). Measurements (1) and (2) both led accurately to the expected result above (they are anti-correlated and equal signals), which were used to calibrate measurements (3). An electron temperature of $\sim$10 mK was deduced from these measurements[22].

We were ready at this point to measure the two-particle cross-correlation. All four QPCs were tuned to $T_{QPC} = 0.5$ while the MDG was left open, hence, turning the two MZIs into a single two-particle interferometer. Equal DC voltages were applied to sources S1 and S2 with two separated power supplies $V_{S1} = V_{S2} = 7.8\,\mu$V ($I_{S1} = I_{S2} \equiv I = 0.3$ nA). For that voltage, there is at most a single electron in each of the four trajectories of the interferometer (the wave packet's width, 15–30 μm, estimated from the current and the estimated drift velocity (~ 3–6×10^6 cm s^{-1}), is bigger then the interferometer's path length, being $\sim$8 μm). This guaranteed a stronger overlap between the wavefunctions of the two electrons, and minimized Coulomb interaction among the electrons (thus eliminating nonlinear effects in the interferometer[19]). The measured fluctuations in D2 and D4 were averaged over some 30,000 electrons, amplified by two separate amplification channels (each fed by its own power supply), and finally cross-correlated. In order to verify

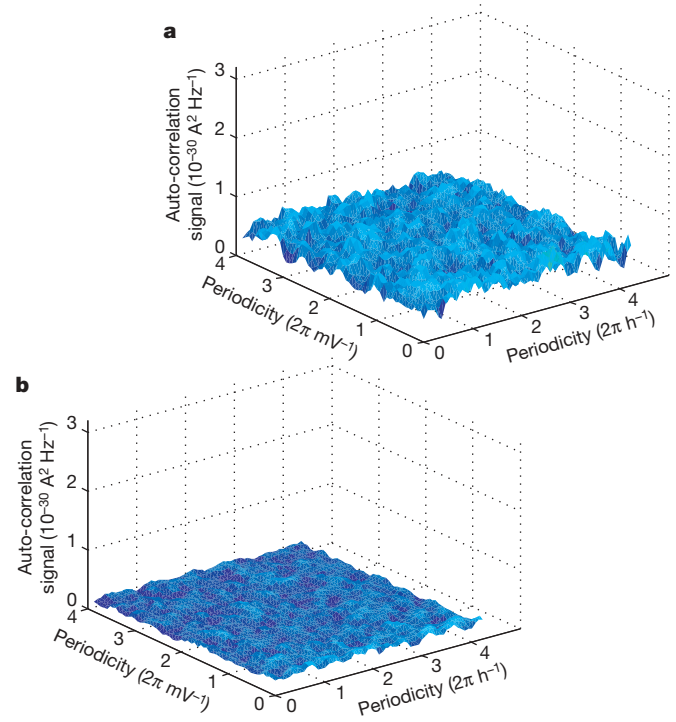


Figure 3 | Analysis and two-dimensional FFT of auto-correlation (shot noise) for an open 'middle gate'. Panels **a** and **b** show two-dimensional FFTs of shot noise measurements in D2 and D4, respectively. The noise is totally featureless, with no sign of Aharonov–Bohm oscillations above the background.

flux insensitivity in each drain separately, we first measured the shot noise in D2 and in D4 as function of the magnetic flux (varying V_{MG} and magnetic field). The noise, with a spectral density of $S = 0.5eI \approx 2.4 \times 10^{-29}\,A^2\,Hz^{-1}$, was found to be featureless. For further assurance, a two-dimensional fast Fourier transform (FFT) of the measurements was calculated, with the results shown in Fig. 3a and b. Again, the transforms were without any feature above our measurement resolution of $\sim 2 \times 10^{-31}\,A^2\,Hz^{-1}$, confirming the absence of flux periodicity in the noise (as was found also in the transmission).

We estimate now the expected magnitude of the cross-correlation signal from equation (3). When $\Phi_{total} = 0$, a maximum anti-correlation signal of the current fluctuations at the drains $S_{D2D4} = \langle \Delta I_{D2} \cdot \Delta I_{D4} \rangle$ is expected. It can be shown that the expected value of the cross-correlation spectral density, for a 100% visibility, is the same as that of the noise of a single QPC, that is, $S_{QPC} = 2eIT_{QPC}(1 - T_{QPC})$, or $0.5eI$ (for $T_{QPC} = 0.5$). As for $\Phi_{total} = \pi$ the cross-correlation signal is expected to vanish, we may conclude that the cross-correlation signal should oscillate with Φ_{total}, $S_{D2D4} = -0.25eI(1 - \sin \Phi_{total})$, with amplitude $1.2 \times 10^{-29}\,A^2\,Hz^{-1}$ for $I = 0.3$ nA.

Without currents in the sources, the cross-correlation signal was featureless (the background), with an average over the two-dimensional FFT of $\sim 2 \times 10^{-31}\,A^2\,Hz^{-1}$ (not shown). The cross-correlation measurement with $I = 0.3$ nA is shown in Fig. 4. The Aharonov–Bohm oscillations are already visible in the raw data (Fig. 4a bottom panel). In the two-dimensional FFT (Fig. 4b), one sees a sharp peak corresponding to a period of 0.58 mV in V_{MG} (with the same voltage applied to MG1 and MG2) and a period of 42.5 min in time (being proportional to the magnetic field decay). The square root of the integrated power under the FFT peak (the amplitude of the Aharonov–Bohm oscillations) is $3.0 \times 10^{-30}\,A^2\,Hz^{-1}$. A roughly similar magnitude was observed also at a bulk filling factor of 2. Moreover, we could directly resolve the Aharonov–Bohm oscillations as a function of V_{MG} and time separately by coherent time averaging. As the magnetic field decayed in time, thus adding continuously an Aharonov–Bohm phase, this extra phase could be compensated for by shifting subsequent scans in V_{MG} according to the decay rate found in the two-dimensional FFT, leading to the negative oscillatory cross-correlation fringes shown in the top left panel of Fig. 4a. Similarly, the oscillations as a function of magnetic field have been extracted (top right panel, Fig. 4a). In Fig. 4c we provide the vector representation of the periodicities (inverse of periods) of each individual MZI (from Fig. 2) and that of the two-particle interferometer, the last being, quite accurately, the sum of the two. This is expected, as the rate of change of the Aharonov–Bohm flux

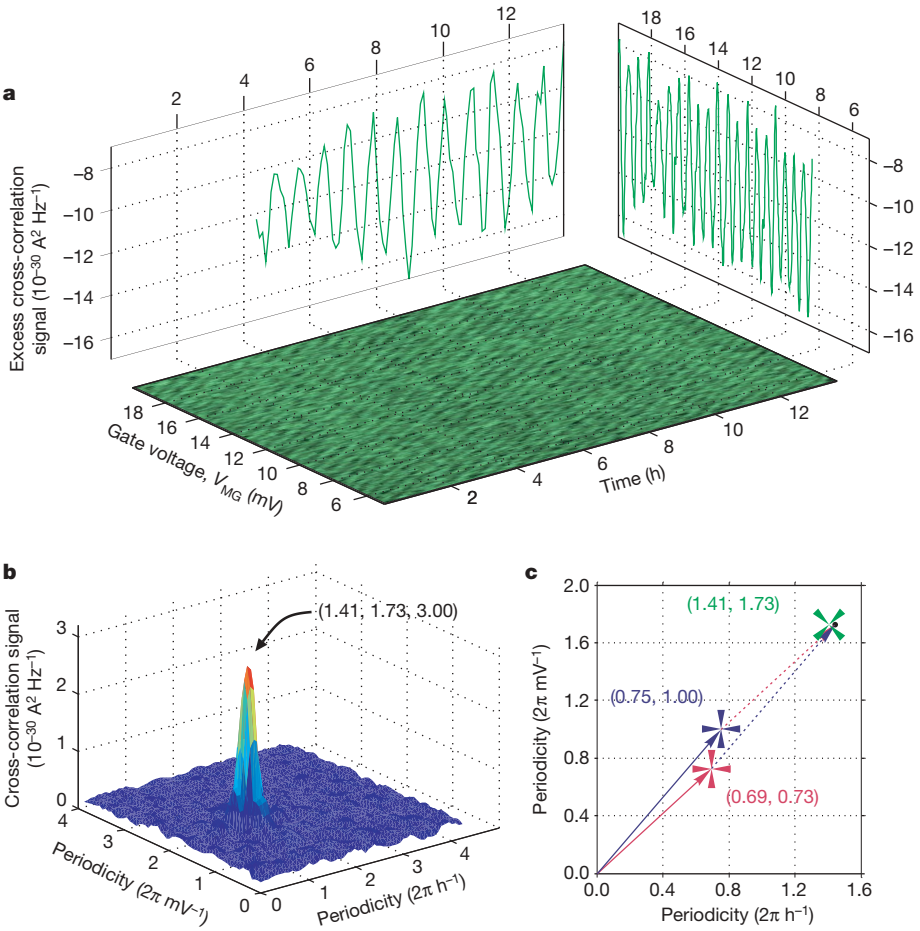


Figure 4 | Cross-correlation of the current fluctuations in D2 and D4. **a**, Bottom, two-dimensional colour plot of the raw data as function of V_{MG} and time (magnetic field). The periodicity is already visible in the raw data. Top right panel, coherent averaging of some 50 traces as function of V_{MG}, by correcting for the added phase due to the decaying magnetic field (see text). Strong Aharonov–Bohm oscillations are seen in the negative excess cross-correlation (the part of the cross-correlation above the background, resulting from an injected current of 0.3 nA at each source). Note that the mean non-oscillating part of the excess cross-correlation is $-1.2 \times 10^{29}\,A^2\,Hz^{-1}$, as expected. Top left panel, similar averaging of the data but at a fixed V_{MG}. The somewhat different visibilities in both panels are due to analysis that must be done in different regions of the two-dimensional plot. **b**, Two-dimensional FFT of the cross-correlation signal. A strong peak is visible, with an integrated power $3.0 \times 10^{-30}\,A^2\,Hz^{-1}$. **c**, A vector representation of the different periodicities. The two vectors starting from the origin and ending at the blue and red crosses are the two-dimensional periodicities of the two MZIs. The green cross is the two-dimensional periodicity of the cross-correlation signal of the two-particle interferometer. The vectorial sum of the periodicities of the two MZIs (black dot) agrees excellently with the corresponding two-dimensional periodicity of the two-particle interferometer.

of the two-particle interferometer is the sum of the rates of the two MZIs.

Compared with the expected amplitude of the cross-correlation oscillations, $1.2 \times 10^{-29}\,A^2\,Hz^{-1}$, we measured an amplitude of $3.0 \times 10^{-30}\,A^2\,Hz^{-1}$. Our results are reasonably accurate, as the measurements have been repeated a few times and over long periods of integration times, lowering the uncertainty to below $10^{-31}\,A^2\,Hz^{-1}$. At least two factors could lead to the lower cross-correlation signal. First, although we have no theory for it, it is likely that the lower visibility in each of the MZI's, v_{MZI1} and v_{MZI2}, will lower the cross-correlation signal by $v_{MZI1} \times v_{MZI2}$. Whereas the visibilities at zero applied d.c. voltage were ~80% (Fig. 2), the visibilities at the applied DC voltage $V_S = 7.8\,\mu V$ were found to be ~70% (ref. 19). Second, our finite temperature (~10 mK) will lower the shot noise by ~22%, affecting the cross-correlation signal similarly. These two effects alone will lower the expected cross-correlation signal to $\sim 4.6 \times 10^{-30}\,A^2\,Hz^{-1}$, which is about 1.5 times higher than the measured one. This discrepancy is still not understood.

Our direct observation of interference between independent particles provides a reliable scheme to entangle separate, but indistinguishable, quantum particles. The present demonstration, done with electrons, reproduces the original Hanbury Brown and Twiss experiments[13,14], which were performed with classical waves. Such experiments are central to the study of the wavefunctions of multiple particles. Our scheme has the potential to test Bell inequalities[2,3,23]; however, taking into account the finite temperature, it seems that the possibility of violating Bell inequalities in our measurements (with a visibility of merely 25%) requires further theoretical analysis.

Received 8 February; accepted 22 May 2007.

1. Feynman, R. P., Leighton, R. B. & Sands, M. *The Feynman Lectures on Physics* Vol. III, *Quantum Mechanics* (Addison-Wesley, New York, 1965).
2. Yurke, B. & Stoler, D. Bell's-inequality experiments using independent-particle sources. *Phys. Rev. A* **46,** 2229–2234 (1992).
3. Samuelsson, P., Sukhorukov, E. V. & Buttiker, M. Two-particle Aharonov-Bohm effect and entanglement in the electronic Hanbury Brown-Twiss setup. *Phys. Rev. Lett.* **92,** 026805 (2004).
4. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47,** 777–780 (1935).
5. Mandel, L. Quantum effects in one-photon and two-photon interference. *Rev. Mod. Phys.* **71,** S274–S283 (1999).
6. Kltenbaek, R. *et al.* Experimental interference of independent photons. *Phys. Rev. Lett.* **96,** 240502 (2006).
7. Kumar, A. *et al.* Experimental test of the quantum shot noise reduction theory. *Phys. Rev. Lett.* **76,** 2778–2781 (1996).
8. Oliver, W. D., Kim, J., Liu, R. C. & Yamamoto, Y. Hanbury Brown and Twiss-type experiment with electrons. *Science* **284,** 299–301 (1999).
9. Henny, M. *et al.* The fermionic Hanbury Brown and Twiss experiment. *Science* **284,** 296–298 (1999).
10. Klessel, H., Renz, A. & Hasselbach, F. Observation of Hanbury Brown-Twiss anticorrelations for free electrons. *Nature* **418,** 392–394 (2002).
11. Reznikov, M., Heiblum, M., Shtrikman, H. & Ma'halu, D. Temporal correlation of electrons: Suppression of shot noise in a ballistic quantum point contact. *Phys. Rev. Lett.* **75,** 3340–3343 (1995).
12. Gavish, U., Levinson, Y. & Imry, Y. Shot-noise in transport and beam experiments. *Phys. Rev. Lett.* **87,** 216807 (2001).
13. Hanbury Brown, R. & Twiss, R. Q. A new type of interferometer for use in radio astronomy. *Phil. Mag.* **45,** 663–682 (1954).
14. Hanbury Brown, R. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **177,** 27–29 (1956).
15. Ji, Y. *et al.* An electronic Mach-Zehnder interferometer. *Nature* **422,** 415–418 (2003).
16. Halperin, B. I. Quantized Hall conductance, current-carrying edge states, and the existence of extended states in a two-dimensional disordered potential. *Phys. Rev. B* **25,** 2185–2190 (1982).
17. Lesovik, B. G. Excess quantum shot noise in 2D ballistic point contacts. *JETP Lett.* **49,** 592–594 (1989).
18. Born, M. & Wolf, E. *Principles of Optics* 7th edn 348–352 (Cambridge Univ. Press, Cambridge, UK, 1999).
19. Neder, I. *et al.* Unexpected behavior in a two-path electron interferometer. *Phys. Rev. Lett.* **96,** 016804 (2006).
20. Neder, I. *et al.* Entanglement, dephasing and phase recovery via cross-correlation measurements of electrons. *Phys. Rev. Lett.* **98,** 036803 (2007).
21. Aharonov, Y. & Bohm, D. Significance of electromagnetic potentials in quantum theory. *Phys. Rev.* **115,** 485–491 (1959).
22. Heiblum, M. Quantum shot noise in edge channels. *Phys. Status Solidi B* **243,** 3604–3616 (2006).
23. Bell, J. S. On the Einstein, Podolsky, Rosen paradox. *Physics* **1,** 195–200 (1964).

Acknowledgements We thank Y. Imry, U. Gavish, M. Buttiker, P. Samuelsson and D. Rohrlich for discussions. The work was partly supported by the Israeli Science Foundation (ISF), the Minerva foundation, the German Israeli Foundation (GIF), the German Israeli Project cooperation (DIP), and the Ministry of Science - Korea Program. Y.C. was supported by the Korea Research Institute of Standards and Science (KRISS), the Korea Foundation for International Cooperation of Science and Technology (KICOS), the Nanoscopia Center of Excellence at Hanyang University through a grant provided by the Korean Ministry of Science and Technology, and by the Priority Research Centers Program funded by the Korea Research Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.H. (heiblum@wisemail.weizmann.ac.il).

LETTERS

A reversible wet/dry adhesive inspired by mussels and geckos

Haeshin Lee[1], Bruce P. Lee[4] & Phillip B. Messersmith[1,2,3]

The adhesive strategy of the gecko relies on foot pads composed of specialized keratinous foot-hairs called setae, which are subdivided into terminal spatulae of approximately 200 nm (ref. 1). Contact between the gecko foot and an opposing surface generates adhesive forces that are sufficient to allow the gecko to cling onto vertical and even inverted surfaces. Although strong, the adhesion is temporary, permitting rapid detachment and reattachment of the gecko foot during locomotion. Researchers have attempted to capture these properties of gecko adhesive in synthetic mimics with nanoscale surface features reminiscent of setae[2–7]; however, maintenance of adhesive performance over many cycles has been elusive[2,8], and gecko adhesion is greatly diminished upon full immersion in water[9,10]. Here we report a hybrid biologically inspired adhesive consisting of an array of nanofabricated polymer pillars coated with a thin layer of a synthetic polymer that mimics the wet adhesive proteins found in mussel holdfasts. Wet adhesion of the nanostructured polymer pillar arrays increased nearly 15-fold when coated with mussel-mimetic polymer. The system maintains its adhesive performance for over a thousand contact cycles in both dry and wet environments. This hybrid adhesive, which combines the salient design elements of both gecko and mussel adhesives, should be useful for reversible attachment to a variety of surfaces in any environment.

The adhesive forces of the gecko have been observed to be on the order of 40 μN or more per seta[11,12] and 10 nN per spatula[13]. Gecko adhesion has been explained as arising from weak secondary bond forces such as van der Waals[11]. However, adhesion of a single spatula varies as a function of humidity and is dramatically reduced under water[9,10], suggesting some contribution from capillary forces. Contact mechanics arguments have been invoked to explain the subdivision of the setal contact surface into multiple independent nanosized spatulae, giving rise to enhancement of the mechanical behaviour[14]. Although the scaling depends on contact geometry, for the idealized case of a hemispherical contact, the theory suggests that the adhesion strength scales with $n^{1/2}$, where n is the number of independent contacts into which the area is subdivided. The contact splitting theory qualitatively explains the scaling of dry adhesive systems used by some amphibians and insects, and provides guidance for development and optimization of synthetic gecko mimics[2,6,15,16]. Synthetic gecko adhesives that exhibit dry adhesion have been fabricated from polymers[2–4] as well as multiwalled carbon nanotubes[5]. However, maintenance of adhesion during repetitive contacts has only been demonstrated for a few contact cycles[2,8], and none have been shown to function under water.

A celebrated biological model for wet adhesion is the mussel, which is well known for its ability to cling to wet surfaces[17,18]. Mussels secrete specialized adhesive proteins containing a high content of the catecholic amino acid 3,4-dihydroxy-L-phenylalanine (DOPA)[19–21]. Both natural and synthetic adhesives containing DOPA and its derivatives have demonstrated strong interfacial adhesion strength[22–25]. Using single-molecule measurements in aqueous media, we recently demonstrated that DOPA formed extraordinarily strong yet reversible bonds with surfaces[26]. In fact, the force necessary to dissociate DOPA from an oxide surface (~800 pN) was the highest ever observed for a reversible interaction between a small molecule and a surface[26]. We speculated that the incorporation of mussel-mimetic polymers into a gecko-foot-mimetic nanoadhesive would yield strong yet reversible wet/dry adhesion—a property that existing materials do not exhibit.

Our strategy employed arrays of gecko-mimetic nanoscale pillars coated with a thin mussel-mimetic polymer film (Fig. 1). The designs of both pillar array and coating polymer incorporated our current knowledge of the respective adhesive systems of gecko and mussel. For the pillar array, the primary design criteria include dimensions of

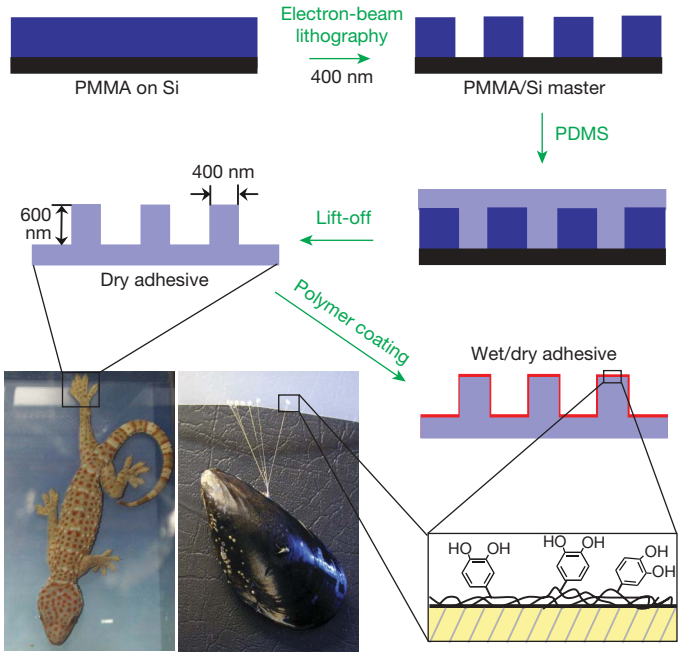


Figure 1 | Rational design and fabrication of wet/dry hybrid nanoadhesive. Electron-beam lithography was used to create an array of holes in a PMMA thin film supported on Si (PMMA/Si master). PDMS casting onto the master is followed by curing, and lift-off resulted in gecko-foot-mimetic nanopillar arrays. Finally, a mussel-adhesive-protein-mimetic polymer is coated onto the fabricated nanopillars. The topmost organic layer contains catechols, a key component of wet adhesive proteins found in mussel holdfasts.

[1]Biomedical Engineering Department, [2]Material Science and Engineering Department, [3]Institute for BioNanotechnology in Medicine, Northwestern University, Evanston, Illinois 60208, USA. [4]Nerites Corporation, 525 Science Drive, Suite 215, Madison, Wisconsin 53711, USA.

the pillars and their spacing, as well as the stiffness of the material[2,15,16]. For flexibility in adapting to rough surfaces, both the supporting substrate and the pillar material were fabricated from poly(dimethylsiloxane) (PDMS) elastomer, which is a well-known organic material with a long history of use in microfabrication[27]. We successfully fabricated (Fig. 1) arrays of PDMS pillars 200, 400 and 600 nm in diameter, with 1–3 μm centre-to-centre distance, and 600–700 nm in height, using electron-beam lithography. The pillar arrays are supported on a continuous film of PDMS (2–3 mm in thickness), with each PDMS pillar representing a single spatula found at the surface of a gecko foot (Fig. 2a, b). Pillar arrays of 400 nm diameter and 600 nm height were tested for adhesion.

Inspection of mussel adhesive protein composition gave insight into a rational design for a mussel-mimetic polymer. First, the synthetic polymer should have a high catechol content since DOPA accounts for as much as 27% of amino acids in the adhesive proteins found at the interface between mussel byssal pads and their substrate[21]. Second, long-lasting waterproof adhesion requires polymers with low water solubility to prevent their loss into the aqueous medium[28]. Thus, we synthesized poly(dopamine methacrylamide-co-methoxyethyl acrylate) (p(DMA-co-MEA); Fig. 2c) through free-radical polymerization where the adhesive monomer, DMA, accounts for 17% of this copolymer by weight (^{1}H-nuclear magnetic resonance spectroscopy). p(DMA-co-MEA) has a high molecular mass and is insoluble in water.

p(DMA-co-MEA) was applied to the PDMS pillar array by dip coating in an ethanol solution of p(DMA-co-MEA). X-ray photoelectron spectroscopy analysis of the coated substrate indicated a thin coating (<20 nm) as demonstrated by the presence of both silicon (103 eV, Si 2*p*) from PDMS and nitrogen (399 eV, N 1*s*) from p(DMA-co-MEA) (Supplementary Fig. 1). A thin coating was desired for minimizing the change in pillar dimensions during coating, which was confirmed by scanning electron microscopy after coating with p(DMA-co-MEA) (Fig. 2d). We refer to the resulting flexible organic nanoadhesive as 'geckel', reflecting the inspiration from both gecko and mussel.

The performance of geckel adhesive was evaluated using an atomic force microscopy (AFM) system fully integrated with optical microscopy, which permitted simultaneous measurement of the adhesive contact force along with clear visualization of nanoscale contact area down to the single pillar level. In a typical adhesion experiment (Fig. 3), the AFM piezo was used to bring a tipless cantilever (Si_3N_4) into contact with the geckel pillar array, and upon retraction the force necessary to separate the cantilever from the pillar array was measured. Furthermore, independently changing the spacing d between pillars ($d = 1$, 2 and 3 μm) and the angle of orientation θ between the pillar array and the cantilever axis (Fig. 3b) allowed us to control the number of pillars contacting the cantilever precisely from one to six. For example, a geckel adhesive with $d = 3$ μm and $\theta = 45^\circ$ resulted in a single pillar contact (Fig. 3c), whereas $d = 1$ μm and $\theta = 0^\circ$ resulted in six pillars interacting with the cantilever simultaneously (Fig. 3d, Supplementary Video 1).

Adhesion experiments were performed both in air and under water for uncoated (hereafter 'gecko') and p(DMA-co-MEA) coated ('geckel') pillar arrays (Fig. 4). Pillar-resolved (that is, area-defined) force measurements showed strong adhesive forces when the cantilever is pulled away from the pillar surface. Figures 4a and b show typical force–distance curves, with each curve representing a specific number (1 to 6) of 400 nm diameter pillars interacting with the Si_3N_4 cantilever surface. The pull-off force was determined from each force–distance curve and mean values from multiple experiments plotted in Fig. 4d as a function of the number of contacting pillars. The observed linear increase in force with pillar number indicates constructive force accumulation, that is, simultaneous detachment of individual pillars from the cantilever. The adhesive force per pillar (Fig. 4e) was calculated from the individual slopes: 39.8 ± 2 nN (gecko in air), 5.9 ± 0.2 nN (gecko in water), 120 ± 6 nN (geckel in air) and 86.3 ± 5 nN (geckel in water).

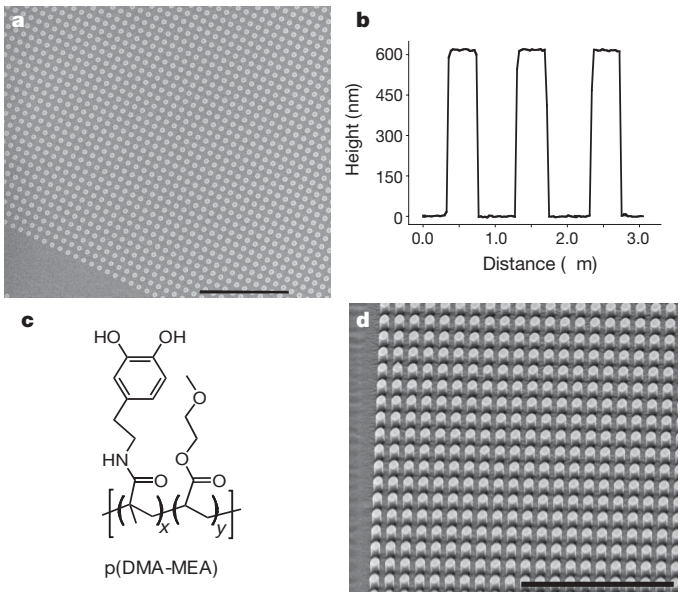


Figure 2 | Fabricated gecko and geckel adhesives. **a**, Scanning electron microscopy image of gecko nanopillar array fabricated using electron-beam lithography. Scale = 10 μm. **b**, AFM linescan of the gecko nanopillars. The height and diameter of the pillars used in this study were 600 and 400 nm, respectively. The apparent widening of the pillars near the base is believed to be an artefact arising from the pyramidal shape of the AFM tip used for imaging. **c**, Chemical structure of the mussel-mimetic polymer, p(DMA-co-MEA), which is applied to the surface of the gecko nanopillars. **d**, Scanning electron microscopy image of geckel adhesive after coating nanopillar array with p(DMA-co-MEA). Scale = 10 μm.

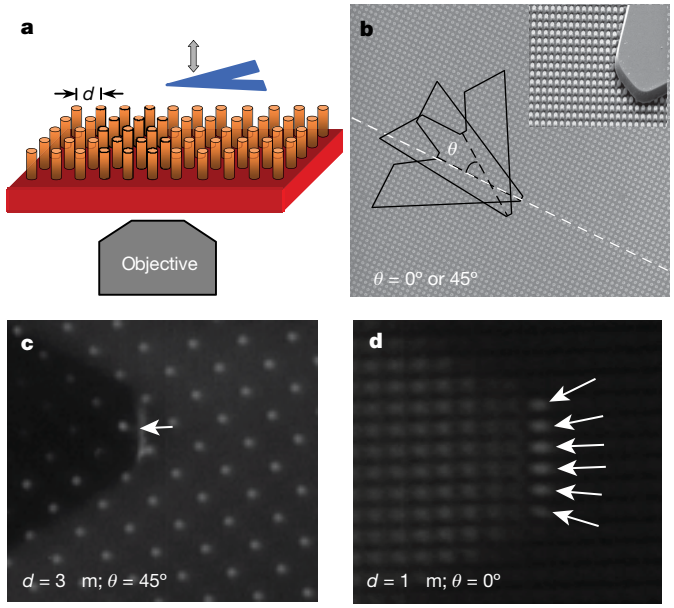


Figure 3 | AFM method for adhesion measurement and imaging of contact area at the single pillar level. **a**, Adhesion was measured by bringing a tipless AFM cantilever into contact with the nanopillar array and then retracting while the contact area is imaged from below. **b**, The number of pillars contacting the cantilever was controlled through the distance d between pillars, and the angle θ between the cantilever and the axis of the pillar array. The inset shows a scanning electron microscopy image of a cantilever contacting a pillar array. **c, d**, Optical microscope images showing one (**c**) and six (**d**) pillar contacts achieved with $d = 3$ μm and $\theta = 45^\circ$, and $d = 1$ μm and $\theta = 0^\circ$, respectively.

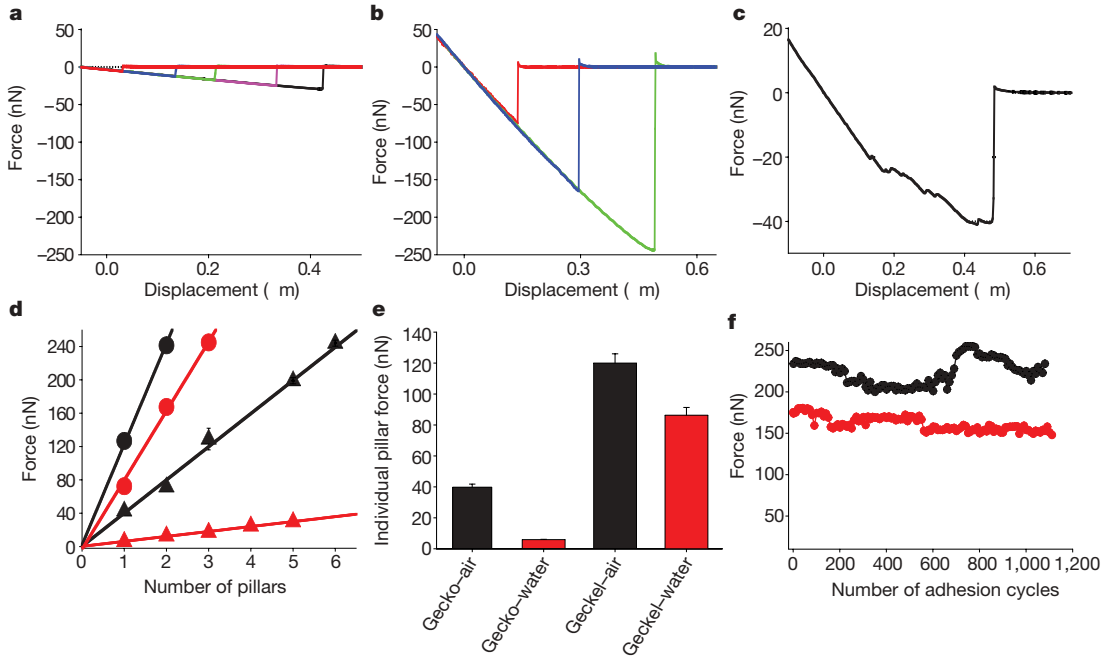


Figure 4 | Force–distance curves and adhesion strength of geckel adhesive. All data are for contact with a Si_3N_4 cantilever. **a**, **b**, Retraction force–distance curves for uncoated (**a**) and p(DMA-co-MEA) coated (**b**) pillars in water. Force–distance curves were obtained for contact with one (red), two (blue), three (green), four (pink), and five (black) pillars. **c**, Retraction force–distance curve for contact between cantilever and flat p(DMA-co-MEA)-coated PDMS (contact area = 5.3 μm^2). **d**, Mean separation force values versus number of pillars for gecko (triangle) and geckel (circle) in water (red) and air (black) ($n > 60$, for each data point). **e**, Adhesion force per pillar, obtained from the slopes of the regression lines shown in **d**. **f**, Performance of geckel adhesive during multiple contact cycles in water (red) and air (black). Error bars represent standard deviation.

Although the addition of p(DMA-co-MEA) coating on the pillars significantly increased dry adhesion, the enhancement of wet adhesion was particularly dramatic, as the wet adhesive force per pillar increased nearly 15 times (from 5.9 to 86.3 nN per pillar, Si_3N_4) when coated with p(DMA-co-MEA). The geckel wet-adhesion strength was also high when tested against other surfaces: titanium oxide (130.7 ± 14.3 nN per pillar) and gold (74.3 ± 4.1 nN per pillar) (Supplementary Fig. 2). The versatility of geckel is not surprising given recent single-molecule force experiments showing the ability of DOPA to interact strongly with both organic and inorganic surfaces[26]. These interactions can take many forms, including metal coordination bonds, pi electron interactions, and covalent bonds. The lower adhesion strength of geckel on gold is in qualitative agreement with our earlier single-molecule pull-off and polymer adsorption studies that indicated DOPA interacts less strongly to gold than to titanium oxide[26,29].

Furthermore, as suggested by our previous study in which we observed the strong bond between DOPA and a metal oxide surface to rupture upon pulling and then re-form when brought back into contact with the surface[26], we speculated that geckel hybrid nanoadhesive may exhibit reversible adhesion to substrates. Repetitive AFM force measurements showed that geckel's wet- and dry-adhesion power was only slightly diminished during many cycles of adhesion, maintaining 85% in wet (red) and 98% in dry (black) conditions after 1,100 contact cycles (Fig. 4f). To our knowledge, no other gecko-mimetic adhesive has demonstrated efficacy for more than a few contact cycles[2,8], and none have been shown to work under water. Control experiments involving pillar arrays coated with the catechol-free polymer p(MEA) showed lower adhesion strength (26 nN per pillar for the first contact cycle) as well as rapid decay in the adhesion performance under cyclic testing (Supplementary Fig. 3), emphasizing the importance of the mussel-mimetic catechol groups in enhancing wet adhesion as well as anchoring the p(DMA-co-MEA) polymer on the pillar array. At the same time, it appears that the nanostructured surface is essential to the observed geckel adhesive behaviour. Force measurements on flat substrates coated with p(DMA-co-MEA) indicated a complex peeling behaviour initiating at low adhesive strength (Fig. 4c), which contrasts with the linear force accumulation behaviour exhibited by the geckel adhesive (Fig. 4d).

The geckel nanoadhesive was shown to be highly effective at adhering reversibly to surfaces under water, and with functional performance resembling that of a sticky note. Although we must be cautious in extrapolating our results to larger areas because of the challenges associated with maintaining equal load sharing among a large number of posts, in its current form (400 nm pillar diameter and 1 μm spacing) a 1 cm^2 surface area of geckel adhesive would transmit 9 N of force under water (90 kPa). It is interesting to note that this value is similar to estimates for the strength of gecko dry adhesion[9,11,12], suggesting that under wet conditions our hybrid geckel adhesive may perform as well as gecko adhesives do under dry conditions. Further refinement of the pillar geometry and spacing, the pillar material, and mussel-mimetic polymer may lead to even greater improvements in performance of this nanostructured adhesive. The results of this study should be of relevance to the design of wet temporary adhesives for medical, industrial, consumer and military settings.

METHODS SUMMARY

For the fabrication of gecko-mimetic adhesive, we first used electron-beam lithography to create a pattern of holes in a poly(methyl methacrylate) (PMMA) film supported on a silicon wafer (negative mould). To create a gecko-mimetic pillar array, sol phase PDMS was cast onto the negative mould, thermally solidified, and then lifted off from the substrate to yield a positive array of PDMS pillars (~400 nm in diameter and 600 nm in height) supported on a continuous PDMS film. Mussel-mimetic polymer, p(DMA-co-MEA), was synthesized by free radical copolymerization of the DMA and MEA monomers, and its molecular weight was analysed by gel permeation chromatography (Wyatt Technology). Finally, the geckel adhesive was prepared by dip-coating PDMS pillar arrays into an ethanol solution of p(DMA-co-MEA) for 3 h. Surface chemical compositions were analysed by X-ray photoelectron spectroscopy (Omicron) and time-of-flight secondary ion mass spectrometry (ToF-SIMS, Physical Electronics). Pillar arrays were imaged by AFM (Veeco) and scanning electron microscopy (FEI). Adhesive forces under dry/wet conditions were

determined by AFM (MFP-1D, Asylum Research) equipped with tipless cantilevers. The contact area between tip and pillar array was precisely controlled by the distance between pillars ($d = 1$, 2 and 3 μm) and the angle between cantilever and pillar axis (θ), and was determined by optical imaging using a 40× objective and fibre-optic illumination.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 November 2006; accepted 30 May 2007.

1. Ruibal, R. & Ernst, V. The structure of the digital setae of lizards. *J. Morphol.* **117,** 271–293 (1965).
2. Geim, A. K. *et al.* Microfabricated adhesive mimicking gecko foot-hair. *Nature Mater.* **2,** 461–463 (2003).
3. Northen, M. T. & Turner, K. L. A batch fabricated biomimetic dry adhesive. *Nanotechnology* **16,** 1159–1166 (2005).
4. Sitti, M. & Fearing, R. Synthetic gecko foot-hair micro/nano-structures as dry adhesives. *J. Adhes. Sci. Technol.* **17,** 1055–1073 (2003).
5. Yurdumakan, B., Raravikar, N. R., Ajayan, P. M. & Dhinojwala, A. Synthetic gecko foot-hairs from multiwalled carbon nanotubes. *Chem. Commun.* **30,** 3799–3801 (2005).
6. Peressadko, A. & Gorb, S. N. When less is more: Experimental evidence for tenacity enhancement by division of contact area. *J. Adhesion* **80,** 1–5 (2004).
7. Crosby, A. J., Hageman, M. & Duncan, A. Controlling polymer adhesion with "Pancakes". *Langmuir* **21,** 11738–11743 (2005).
8. Northen, M. T. & Turner, K. L. Meso-scale adhesion testing of integrated micro- and nano-scale structures. *Sensors Actuators A* **130–131,** 583–587 (2006).
9. Huber, G. *et al.* Evidence for capillary contributions to gecko adhesion from single spatula nanomechanical measurements. *Proc. Natl Acad. Sci. USA* **102,** 16293–16296 (2005).
10. Sun, W., Neuzil, P., Kustandi, T. S., Oh, S. & Samper, V. D. The nature of the gecko lizard adhesive force. *Biophys. J.* **89,** L14–L16 (2005).
11. Autumn, K. *et al.* Evidence for van der Waals adhesion in gecko setae. *Proc. Natl Acad. Sci. USA* **99,** 12252–12256 (2002).
12. Autumn, K. *et al.* Adhesive force of a single gecko foot-hair. *Nature* **405,** 681–685 (2000).
13. Huber, G., Gorb, S. N., Spolenak, R. & Arzt, E. Resolving the nanoscale adhesion of individual gecko spatulae by atomic force microscopy. *Biol. Lett.* **1,** 2–4 (2005).
14. Arzt, E., Gorb, S. & Spolenak, R. From micro to nano contacts in biological attachment devices. *Proc. Natl Acad. Sci. USA* **100,** 10603–10606 (2003).
15. Arzt, E. Biological and artificial attachment devices: Lessons for materials scientists from flies and geckos. *Mater. Sci. Engin. C* **26,** 1245–1250 (2006).
16. Spolenak, R., Gorb, S. & Arzt, E. Adhesion design maps for bio-inspired attachment systems. *Acta Biomater.* **1,** 5–13 (2005).
17. Waite, J. H. Nature's underwater adhesive specialist. *Chemtech* **17,** 692–697 (1987).
18. Waite, J. H. Adhesion a la moule. *Integr. Comp. Biol.* **42,** 1172–1180 (2002).
19. Waite, J. H. & Tanzer, M. L. Polyphenolic substance of *Mytilus edulis*: novel adhesive containing L-dopa and hydroxyproline. *Science* **212,** 1038–1040 (1981).
20. Papov, V. V., Diamond, T. V., Biemann, K. & Waite, J. H. Hydroxyarginine-containing polyphenolic proteins in the adhesive plaques of the marine mussel *Mytilus edulis*. *J. Biol. Chem.* **270,** 20183–20192 (1995).
21. Waite, J. H. & Qin, X. X. Polyphenolic phosphoprotein from the adhesive pads of the common mussel. *Biochemistry* **40,** 2887–2893 (2001).
22. Yu, M. & Deming, T. J. Synthetic polypeptide mimics of marine adhesives. *Macromolecules* **31,** 4739–4745 (1998).
23. Frank, B. P. & Belfort, G. Adhesion of *Mytilus edulis* foot protein 1 on silica: ionic effects on biofouling. *Biotechnol. Prog.* **18,** 580–586 (2002).
24. Hwang, D. S., Yoo, H. J., Jun, J. H., Moon, W. K. & Cha, H. J. Expression of functional recombinant mussel adhesive protein Mgfp-5 in *Escherichia coli*. *Appl. Environ. Microbiol.* **70,** 3352–3359 (2004).
25. Lee, B. *et al.* Rapid gel formation and adhesion in photocurable and biodegradable block copolymers with high DOPA content. *Macromolecules* **39,** 1740–1748 (2006).
26. Lee, H., Scherer, N. F. & Messersmith, P. B. Single molecule mechanics of mussel adhesion. *Proc. Natl Acad. Sci. USA* **103,** 12999–13003 (2006).
27. Whitesides, G. M. The origins and the future of microfluidics. *Nature* **442,** 368–373 (2006).
28. Waite, J. H., Andersen, N. H., Jewhurst, S. & Sun, C. Mussel adhesion: finding the tricks worth mimicking. *J. Adhesion* **81,** 1–21 (2005).
29. Dalsin, D. L., Hu, B.-H., Lee, B. P. & Messersmith, P. B. Mussel adhesive protein mimetic polymers for the preparation of nonfouling surfaces. *J. Am. Chem. Soc.* **125,** 4253–4258 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the NIH and NASA for providing funding for this work. We thank J. Jureller and W. Russin for advice on optical imaging, B. Meyer for electron-beam lithography discussions, and V. Dravid and K. Shull for critical reading of the manuscript. Portions of this work used the NUANCE (EPIC, KECK-II and NIFTI) and biological imaging facilities at Northwestern University, the Nanobio facility at the University of Chicago, and the National Magnetic Resonance Facility at the University of Wisconsin-Madison.

Author Contributions P.B.M. planned the project, designed experiments, analysed data and wrote the manuscript. H.L. designed and performed experiments, analysed data and wrote the manuscript. B.P.L. designed and synthesized the polymer and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.B.M. (philm@northwestern.edu).

METHODS

Synthesis of DMA. 20 g of sodium borate and 8 g of $NaHCO_3$ were dissolved in 200 ml of deionized water and bubbled with Ar for 20 min. 10 g of dopamine-HCl (52.8 mmol) was then added, followed by the dropwise addition of 9.4 ml of methacrylate anhydride (58.1 mmol) in 50 ml of THF, during which the pH of solution was kept above 8 with addition of 1 M NaOH as necessary. The reaction mixture was stirred overnight at room temperature with Ar bubbling. The aqueous mixture was washed twice with 100 ml of ethyl acetate and then the pH of the aqueous solution was reduced to less than 2 and extracted with 100 ml of ethyl acetate three times. The final three ethyl acetate layers were combined and dried over $MgSO_4$ to reduce the volume to around 50 ml. 450 ml of hexane was added with vigorous stirring and the suspension was held at 4 °C overnight. The product was recrystallized from hexane and dried to yield 9.1 g of grey solid. ^{1}H-nuclear magnetic resonance spectroscopy (400 MHz, DMSO-*d*/TMS): δ 6.64–6.57 (m, 2H, $C_6HH_2(OH)_2$-), 6.42 (d, 1H, $C_6H_2H(OH)_2$-), 5.61 (s, 1H, -C(=O)-C(-CH_3)=C*H*H), 5.30 (s, 1H, -C(=O)-C(-CH_3)=C*H*H), 3.21 (m, 2H, $C_6H_3(OH)_2$-CH_2-CH_2(NH)-C(=O)-), 2.55 (t, 2H, $C_6H_3(OH)_2$-CH_2-CH_2(NH)-C(=O)-), 1.84 (s, 3H, -C(=O)-C(-CH_3)=CH_2). ^{13}C-nuclear magnetic resonance spectroscopy (400 MHz, DMSO-*d*/TMS): δ 167.3 (s, 1C, -NH-*C*(=O)-C(CH_3)=CH_2), 145.0 (s, 1C, -NH-C(=O)-*C*(CH_3)=CH_2), 143.5–115.5 (6C, C_6H_3(O-C(=O)-CH_3)$_2$), 130.3 (s, 1C, -NH-C(=O)-C(CH_3)=*C*H_2), 41.0 (s, 1C, $C_6H_3(OH)_2$-CH_2-*C*H_2(NH)-C(=O)-), 34.6 (s, 1C, $C_6H_3(OH)_2$-*C*H_2-CH_2(NH)-C(=O)-), 18.7 (s, 1C, -C(=O)-C(-CH_3)=CH_2). Italics indicate the atom yielding the peak.

Synthesis of p(DMA-co-MEA). 12.5 ml of MEA was passed through a column packed with 30 g of Al_2O_3 to remove inhibitor. 7.5 g of purified MEA (57.9 mmol), 1.7 g of DMA (7.4 mmol), and 106 mg of azobisisobutyronitrile (0.64 mmol) were added to 20 ml of DMF in an airtight flask. The solution mixture was degassed through pump–freeze–thaw cycles three times. While sealed under vacuum, the solution was heated to 60 °C and stirred overnight. The reaction mixture was diluted with 50 ml of methanol and added to 400 ml of Et_2O to precipitate the polymer. After precipitating in DCM/Et_2O two more times and drying in the vacuum desiccator, 5.7 g of white, sticky solid was obtained. ^{1}H-nuclear magnetic resonance spectroscopy (400 MHz, $CDCl_3$/TMS): δ 6.81–6.70 (d, br, 2H, $C_6HH_2(OH)_2$-), 6.58 (s, br, 1H, $C_6H_2H(OH)_2$-), 4.20 (s, br, 2H, CH_3-O-CH_2-CH_2-O-C(=O)-), 3.57 (s, br, 2H, CH_3-O-CH_2-CH_2-O-C(=O)-), 3.36 (s, br, 3H, CH_3-OCH_2-CH_2-O-C(=O)-), 2.69 (s, br, 2H, $C_6H_3(OH)_2$-CH_2-CH_2(NH)-C(=O)-), 2.39 (s, br, 1H, -O-C(=O)-C*H*(CH_2-)-CH_2-), 2.14 (s, br, 2H, $C_6H_3(OH)_2$-CH_2-CH_2(NH)-C(=O)-), 1.93 (s, 3H, -NH-C(=O)-C(CH_3)(CH_2-)-CH_2-), 1.68 (m, br, -O-C(=O)-CH(CH_2-)-CH_2-), 0.98 (m, br, -NH-C(=O)-C(CH_3)(CH_2-)-CH_2-). Analysis indicated a 1:6 molar ratio of DMA to MEA in the copolymer. Gel permeation chromatography in concert with multi-angle laser light scattering (Wyatt Technology), with mobile phase of 20 mM LiBr in DMF and Shodex-OH Pak columns: weight-average molecular mass = 252 kDa, polydispersity = 1.73. For control experiments, a catechol-free p(MEA) homopolymer (molecular mass (average) = 100 kDa, Scientific Polymer Products) was used.

Electron-beam lithography. Electron-beam resist (950 PMMA A3, MicroChem) was spin-coated (4,000 r.p.m., 40 s) on silicon wafer several times until the resist thickness, as measured by ellipsometry (Woolam), reached 600–700 nm. The resist was patterned at 30 kV with an area dose of 650–800 μC cm^{-2} using Quanta 600F (FEI). Resist development was performed for 1 min with a solution of methyl isobutyl ketone/isopropanol (1/3, v/v), followed by rinsing with water. The patterned substrates were treated with oxygen plasma (Harrick) for 30 s and repeated 2–3 times to remove residual resist completely from the exposed Si regions. The patterned substrates were then exposed to a triethoxyoctylsilane vapour for 30 min. PDMS was prepared as follows: 4 μl of Pt-catalyst (platinum-divinyl tetramethyl-disiloxane in xylene) and 4 μl of modulator (2,4,6,8-tetramethyl-2,4,6,8-tetravinylcyclotetrasioxane) were added to a 7–8% vinylmethylsiloxane solution (3.5 g). The solution was subsequently mixed with a 25–30% methylhydrosiloxane (1 g) solution. Finally the solution was cured (80 °C) after spin-coating (1,000 r.p.m. for 1 min) onto the PMMA/Si master. The spin-coated substrate was covered either by thin cover glass for force measurements or sylgard-184 PDMS for other experiments such as optical imaging or xray photoelectron spectroscopy. Gecko adhesive was obtained by PDMS pattern lift-off and brief exposure to oxygen plasma (100 W, 30 s) and used within 2–3 h after plasma treatment. Geckel adhesive was prepared by dip-coating gecko adhesive in a 1 mg ml^{-1} solution of p(DMA-co-MEA) in ethanol at 70 °C. Unstructured controls were fabricated in the same manner using flat PDMS, whereas structured controls were fabricated by dip-coating gecko adhesive in p(MEA) using the method described above.

X-ray photoelectron spectroscopy. The presence of p(DMA-co-MEA) and p(MEA) on PDMS surfaces was confirmed by X-ray photoelectron spectroscopy (Omicron) equipped with a monochromatic Al Kα (1,486.8 eV) 300 W X-ray source and an electron gun to eliminate charge build-up.

Atomic force and optical microscopy. All force data were collected on an Asylum Mfp-1D AFM instrument (Asylum Research) installed on a Nikon TE2000 microscope. Spring constants of individual cantilevers (Veecoprobes, NP-20 tipless Si_3N_4 tips) were calibrated by applying the equipartition theorem to the thermal noise spectrum[30]. Owing to the large forces exhibited by the adhesive, only tips exhibiting high spring constants (280–370 pN nm^{-1}) were used. Metal and metal-oxide-coated cantilevers were formed by sputter coating ~10 nm of Au or Ti (a native oxide formed at the Ti surface, TiO_x) using a Denton Vacuum Desk III. The surface composition of each cantilever was confirmed by time-of-flight secondary ion mass spectrometry, using a PHI-TRIFT III (Ga+, 15 keV, Physical Electronics). Cantilevers were treated by oxygen plasma (100 W, 150 mTorr) for 3 min before use. Force measurements were conducted either in millipore water or ambient (air) conditions at a cantilever pulling speed of 2 μm s^{-1}. In wet experiments, optical microscopic examination of the contact region indicated the absence of air bubbles trapped between nanopillars and on the nanopillar surface (not shown). Tapping-mode AFM images were obtained using a multimode Veeco Digital Instrument with a Si cantilever (resonance frequency of 230–280 kHz). Contact area was imaged by an inverted optical microscope using a 40× objective illuminated by a fibre-optic white light source perpendicular to the objective.

30. Hutter, J. L. & Bechhoefer, J. Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* **64**, 1868–1873 (1993).

Vol 448|19 July 2007|doi:10.1038/nature06018

LETTERS

Catastrophic flooding origin of shelf valley systems in the English Channel

Sanjeev Gupta[1], Jenny S. Collier[1], Andy Palmer-Felgate[1] & Graeme Potter[2]

Megaflood events involving sudden discharges of exceptionally large volumes of water are rare, but can significantly affect landscape evolution, continental-scale drainage patterns and climate change[1]. It has been proposed that a significant flood event eroded a network of large ancient valleys on the floor of the English Channel—the narrow seaway between England and France[2–4]. This hypothesis has remained untested through lack of direct evidence, and alternative non-catastrophist ideas have been entertained for valley formation[5,6]. Here we analyse a new regional bathymetric map of part of the English Channel derived from high-resolution sonar data, which shows the morphology of the valley in unprecedented detail. We observe a large bedrock-floored valley that contains a distinct assemblage of landforms, including streamlined islands and longitudinal erosional grooves, which are indicative of large-scale subaerial erosion by high-magnitude water discharges. Our observations support the megaflood model, in which breaching of a rock dam at the Dover Strait instigated catastrophic drainage of a large pro-glacial lake in the southern North Sea basin[2]. We suggest that megaflooding provides an explanation for the permanent isolation of Britain from mainland Europe during interglacial high-sea-level stands[7], and consequently for patterns of early human colonisation of Britain together with the large-scale reorganization of palaeodrainage in northwest Europe[4].

The geographic isolation of Britain from continental Europe is a consequence of high interglacial sea levels that led to marine flooding of the shallow shelf areas of the English Channel and the North Sea[7]. Before the formation of the Dover Strait, however, Britain remained connected to Europe by means of a structural ridge, the Weald–Artois anticline, which extends from southeast England to northwest France (Fig. 1). During interglacial high-sea-level stands, this Chalk ridge formed a narrow isthmus separating marine embayments to the north (North Sea) and to the southwest (English Channel)[4]. To form 'island' Britain it was thus necessary to breach the Weald–Artois ridge; however, the mechanism of the breach remains speculative[4].

Early marine geophysical investigations of the English Channel revealed a ~400-km-long network of submerged and partially infilled valleys carved into the bedrock floor of the shelf[8–10] (Fig. 1). This network extends westwards from the Strait of Dover, collecting the drainages of southern England and northern France before eventually amalgamating to form a single prominent valley, the Hurd Deep[2,8,10]. Formation of the network was explained by a variety of mechanisms: fluvial erosion in response to late Quaternary sea-level lowering[9], glacial erosion[5,6], tidal scouring[11] and erosion by catastrophic flooding[2–4]. However, testing these competing hypotheses has awaited detailed mapping of the sea floor. Here we analyse a regional high-resolution bathymetric grid of the north-central

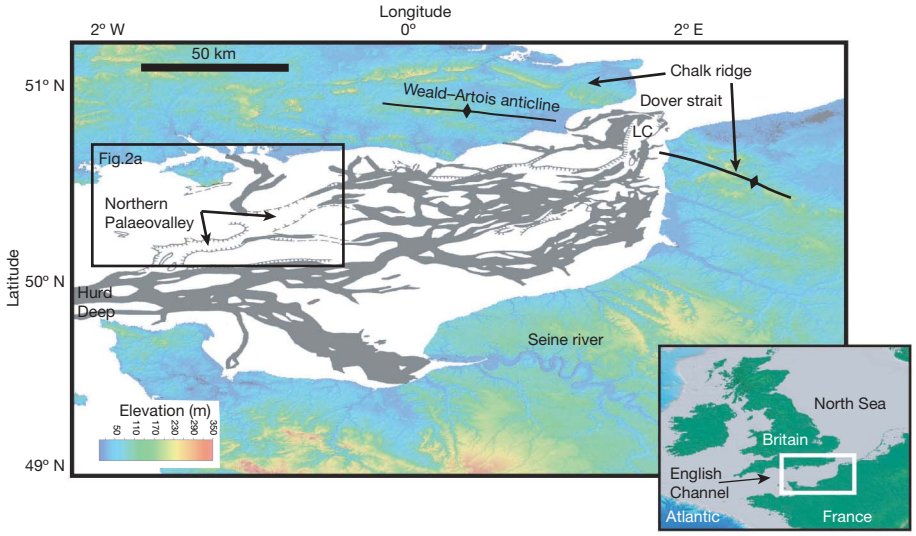


Figure 1 | Location map and inferred distribution of palaeovalleys on the English Channel shelf. Grey indicates valleys filled with sediment; white and hatched, unfilled valleys. The box shows the location of the segment of Northern Palaeovalley that we studied. Onshore topography is shown as a coloured and shaded relief image; data are derived from the NASA Shuttle Radar Topography Mission elevation model. There is a prominent topographic escarpment formed by the Weald–Artois anticline that extends from southeastern England into northwestern France. The Lobourg Channel (LC) in the Dover Strait extends westward into the Northern Palaeovalley. The inset shows the location of the study area with respect to northwest Europe. The map of palaeovalleys is reproduced with permission from ref. 2.

[1]Department of Earth Science and Engineering, Imperial College London, London SW7 2AZ, UK. [2]UK Hydrographic Office, Admiralty Way, Taunton, Somerset TA1 2DN, UK.

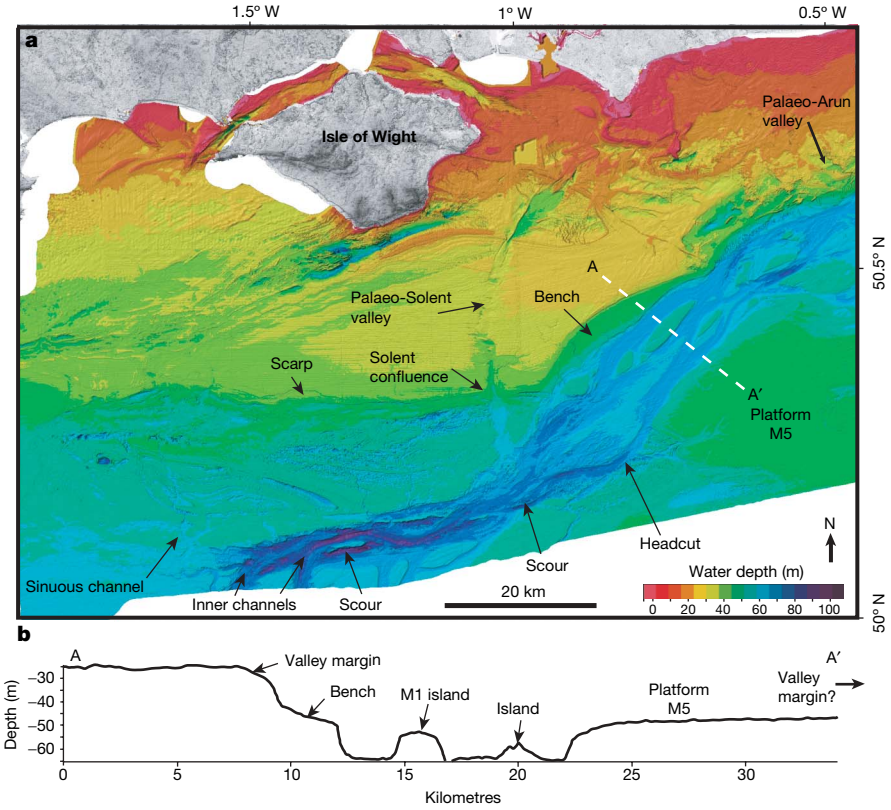


Figure 2 | Sonar bathymetry of the north-central English Channel shelf. **a**, Coloured and shaded relief bathymetry map. Onshore topography is shown as black and white shaded relief. A headcut is a small cataract at the upstream termination of inner channels. Scours are elongate hollows eroded into the channel floor. The scarp is an east–west-trending escarpment defining the northern limit of the palaeovalley. The white dashed line shows the location of the bathymetric profile A–A′. The east–west and east northeast–west southwest striping (closely spaced lines) in the image is an artefact of survey vessel tracks. **b**, Bathymetric profile across the Northern Palaeovalley showing valley margin, bedrock bench and streamlined islands.

English Channel shelf (see Methods). The data show a collection of landforms that, taken together, indicate a catastrophic flood origin.

The bathymetry shows a prominent ~100-km-long, northeast–southwest-trending, linear valley that is eroded into the gently southward-sloping shelf (Fig. 2a). This feature, called the Northern Palaeovalley[11], represents the northern branch of the Channel valley system (Fig. 1) and forms a bedrock-floored valley that is largely devoid of sediment infill. Inner channels within the valley show an anabranching planform. In the northeast, the valley is up to 50 m deep with a floor width of <15 km (Fig. 2a). Traced westwards, it narrows to a prominent <10-km-wide and 40-m-deep inner channel bounded to the north by a 12-km-wide bench. Cross-sections across the valley show rectangular profiles, with valley walls of ~2°, flat valley floors and high width-to-depth ratios (Fig. 2b). In the east, valley margins show streamlined edges (see TL in Fig. 3a and b), which are sharp and erosive (Supplementary Fig. 1). The valley cross-cuts a variety of bedrock lithologies[12], ranging from soft Palaeogene rocks to more resistant Chalk bedrock, indicating that lithology does not significantly control valley morphology.

Evidence of irregular scalloping of the walls by mass movement processes, or spur and gully features typical of dissection by normal fluvial processes, is not apparent. A distinct sub-horizontal bench (~4 km wide, ~25 km long), cut into Chalk bedrock, is observed on the northern valley flank (Fig. 3a, b), indicating at least two episodes of downcutting: one to carve the bench and another to incise it further. The southern margin of the initial incision is not observed in the study area (Fig. 2b); accordance of the bedrock bench and platform M5 indicates that the initial valley was a minimum of 45 km in width. The large size, linear trace and anabranching morphology of the Northern Palaeovalley together with the presence of prominent streamlined valley margins is compelling evidence that valley incision was achieved by high-magnitude flood erosion rather than by normal fluvial processes[13,14] (see Supplementary Information). In contrast, the onshore Seine river valley shows a single-thread sinuous course where it is incised into Chalk bedrock (Fig. 1).

The most striking evidence that the Northern Palaeovalley was formed by catastrophic flooding is the presence of kilometre-scale streamlined islands or mesas (M1–M5 in Fig. 3) with characteristic elongate, lenticular to quadrilateral planforms (M2, M3 in Fig. 3b, c).

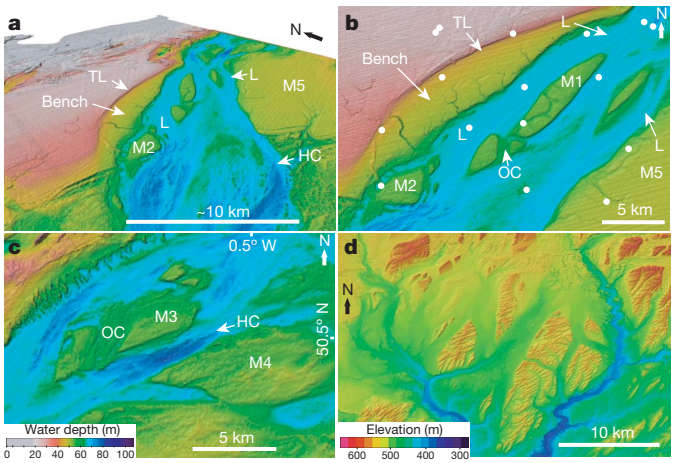


Figure 3 | Details of geomorphology of the Northern Palaeovalley. Location of images indicated in Supplementary Fig. 2. **a**, Three-dimensional perspective view of the Northern Palaeovalley looking northeast. Vertical exaggeration is approximately ×6. Water depth is indicated by the colour scale in **c**. HC, headcut; TL, trim line of valley margin. L, longitudinal lineations. **b**, Vertical view of the northeast reach of the Northern Palaeovalley. The northwestern valley margin shows the presence of a distinct bedrock bench. Prominent streamlined islands are present in the main valley. Flow lines are indicated by longitudinal lineations on the floor of the inner channel. OC, oblique channel. White circles indicate Chalk bedrock from shallow cores. **c**, Vertical view of streamlined islands (M3, M4) in the northeastern reach of the Northern Palaeovalley. There is an oblique channel dissecting island M3. **d**, A coloured and shaded relief image of streamlined islands in the Cheney–Palouse Scabland terrain (Channeled Scabland, Washington). Flood flow direction is to the south. Data are derived from a 10-m-resolution US Geological Survey digital elevation model.

These islands are up to 10 km long and 4 km wide (Fig. 3), with an average length-to-width ratio of 3–4. Shallow coring of the seabed[12] indicates that the islands are eroded into bedrock and do not represent depositional forms (Fig. 3b). The upper surface of the islands is generally smooth and flat, although locally oblique channels cut through small divides at their crest. The streamlined islands bear a striking resemblance to loess islands preserved in the Cheney–Palouse terrain of the Channeled Scabland of Washington, USA[13,14] (Fig. 3d). These formed when outburst flooding from the Pleistocene glacial lake Missoula eroded through loess cover into basalts[13]. The distinctive shape of the islands is thought to be a feature to minimize resistance in fluid flow[15].

Additional evidence for erosion by flooding comes from the presence of longitudinal lineations on the floor of the Northern Palaeovalley (L in Fig. 3a, b). These form alternating ridges and grooves that are oriented parallel to the channel gradient and have long axes sub-parallel to the channel walls (Fig. 3a, b). The grooves are typically <200 m wide, <2–3 m deep and 10–15 km long. They clearly show curvature around intra-channel topography, indicating that they approximate flow streamlines. Similar lineations in the Channeled Scabland are thought to result from erosion by longitudinal vortices developed in high-magnitude flood flows[13]. Also typical of flood erosion are crescent-like scours that taper upstream into V-shaped headcuts (Fig. 3c); we interpreted these to have formed by headward recession of small <10-m-high cataracts.

Southeast of the Isle of Wight, a north–south trending valley (2–4 km wide) is observed that is interpreted as the offshore course of the palaeo-Solent valley[9,11,16]. At its confluence with the Northern Palaeovalley it forms a hanging tributary with a series of south-facing knickpoints (Figs 2a and 4a), suggesting that the Solent was unable to regrade when the Northern Palaeovalley was abruptly incised to its current base level. Another tributary, the 2-km-wide palaeo-Arun (Fig. 2a), has incised a 600-m-wide gap into a prominent east–west-trending bedrock ridge at its confluence with the Northern Palaeovalley (Fig. 4b). Tributary erosion through the bedrock ridge is best explained by an abrupt base-level fall at the confluence caused by rapid incision of the Northern Palaeovalley by catastrophic flooding.

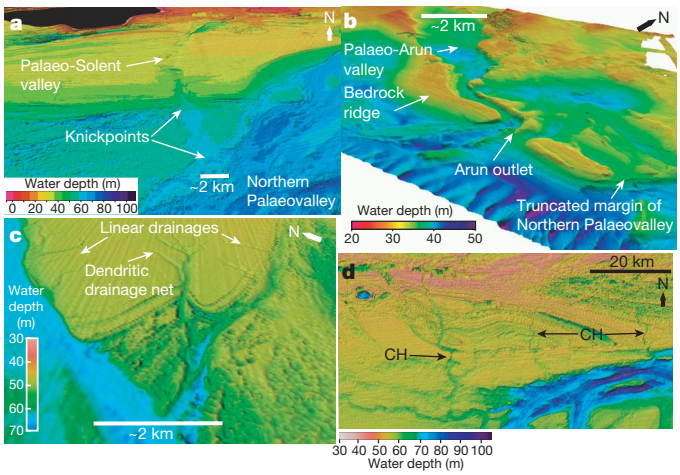


Figure 4 | Bathymetry images showing tributary confluence morphology and post-flooding secondary drainages. **a**, Three-dimensional perspective view of the palaeo-Solent confluence with the Northern Palaeovalley. Distinct knickpoint steps are present at the confluence. Vertical exaggeration is approximately ×6. **b**, Multibeam bathymetry perspective view of the palaeo-Arun confluence with Northern Palaeovalley. Vertical exaggeration is approximately ×6. **c**, Three-dimensional perspective view of post-flooding secondary drainage networks eroded into streamlined islands. **d**, Vertical view of sinuous secondary drainages in the southwestern part of the study area, indicating post-flooding erosion by small rivers. CH, sinuous channels.

Superimposed on the flood-carved topography are a series of dendritic and linear drainages (<250 m wide; <10 m deep) that were carved by normal fluvial erosion processes (Figs 3b and 4c). These debouch to the floor of the Northern Palaeovalley, indicating that they post-date final valley incision. To the southwest of the study area, ~30-km-long sinuous channels reveal evidence of more extensive streams on the English Channel floor (Fig. 4d). Preservation of post-flooding drainages confirms a subaerial setting for valley erosion.

Our study provides the first direct evidence that a megaflood event was responsible for carving the English Channel valley network. Normal fluvial processes cannot explain erosion of the Northern Palaeovalley, because before flooding no significant river was sourced from the Weald–Artois ridge to the east[16] (Fig. 1). Tidal scour is not a viable mechanism because the superposed dendritic drainages indicate subaerial exposure of the valley floor after incision. Erosion by glaciers[5,6] is untenable, because there is no evidence that these advanced into the English Channel[17]. Our observations are consistent with erosion by high-magnitude flood flows, as in the Channeled Scabland, in which analogous landforms were indisputably formed by catastrophic drainage of the glacial lake Missoula[1,13] (see Supplementary Information for additional discussion).

The continuity of the Northern Palaeovalley to the Dover Strait (Fig. 1) indicates an eastern source of floodwaters. Our analysis supports the hypothesis proposed in ref. 2, that catastrophic flooding was caused by overflow of a large pro-glacial lake in the southern North Sea[4]. This lake was impounded by the coalesced Fennoscandian and British ice sheets in the central North Sea and the Weald–Artois barrier across the Dover Strait, and was fed by the Rhine and Thames drainages[16] as well as from melting of the ice sheet itself. Before ice advance in the Elsterian/Anglian stage (conventionally equated to Marine Isotope Stage (MIS) 12), these rivers would have joined on the exposed North Sea shelf and drained northwards[16]. The rise in lake level eventually led to a catastrophic breach of the Weald–Artois structural barrier, with a consequent outburst of floodwaters into the subaerially exposed English Channel[2].

The presence of a bedrock bench at the valley margin indicates that at least two episodes of flooding eroded the valley. The preservation of small truncated and beheaded channels on the upper surface of island M1 (Fig. 3b) is evidence that an interlude of normal fluvial processes operated on the valley floor following the initial episode of flooding but before final valley incision by a second flood episode. We cannot resolve the absolute timing of the flooding events. An Elsterian/Anglian stage (MIS 12) age was initially proposed for overflow of a southern North Sea lake[4]; however, a Saalian/Wolstonian stage (MIS 10–6) age has also been suggested[18]. The Dover Strait was certainly open by the Eemian/Ipswichian stage (MIS 5e)[19]. Our topographic data permit estimation of the water discharges associated with the flooding (see Methods). Maximum peak discharges for the flood events range between $\sim 0.2 \times 10^6\ m^3\ s^{-1}$ and $\sim 1 \times 10^6\ m^3\ s^{-1}$, making them some of the largest megafloods on Earth[20].

Our discovery has several wider implications. The breach of the Weald–Artois anticline reorganized the palaeodrainages of northwest Europe[16] by re-routeing the combined Rhine–Thames river system through the English Channel to form the Channel river, one of the largest palaeodrainages of Europe during late Quaternary low-sea-level stands[21–23]. The breach also led to the permanent separation of Britain from continental Europe during interglacial high-sea-level stands, thus providing an explanation for its recent geographic insularity[7]. These palaeotopographic changes seem to have influenced patterns of early human occupation of Britain. Lower Palaeolithic records of early human activity in southern Britain indicate regular episodes of colonization from Europe[24,25]. Early human populations peaked between the end of MIS 13 and MIS 10, then showed a decline into MIS 8, followed by a sharp drop from MIS 7 (refs 26, 27). From MIS 6 there is a period of human absence of about 100,000 years

(ref. 27). We speculate that flooding-induced changes in topography together with post-flooding diversion of the Rhine–Thames river system created notable barriers to migration across the subaerial Channel floor, thus contributing to the observed pattern of decline and absence. The clear absence of early humans (and also of other mammals such as horses[28]) during MIS 5–4 is a consequence of Britain's resulting isolation from mainland Europe during the last interglacial highstand.

Finally, our results have potential palaeoclimatic significance. It is widely held that massive freshwater pulses into the North Atlantic following outburst flooding from the glacial lake Agassiz caused the Younger Dryas and 8.2 K climatic cooling events through weakening or shutdown of the Atlantic meridional overturning circulation[29,30]. Our discovery of significant flood events in the English Channel that probably discharged into the eastern North Atlantic offers the possibility that outburst flooding from a southern North Sea lake could have forced previous episodes of abrupt climatic cooling.

METHODS SUMMARY

Data acquisition and processing. The bathymetry data were collected for navigational charting purposes by the Maritime and Coastguard Agency and are archived at the United Kingdom Hydrographic Office. The bathymetric soundings were collected during 36 surveys over a 24-year period (1979–2003) with a hull-mounted, single-beam echo-sounder, with individual soundings being converted to depth using measured water sound–velocity profiles. Vessel positioning for the pre-1996 surveys was, in general, by range and bearing from onshore trisponders, and for the post-1996 surveys was by differential global positioning system (GPS). The datasets were generally acquired along transects spaced 62.5 m apart on a bearing of 080° together with 2.5-km-spaced cross-lines. The raw data were hand-edited to remove bad navigational and depth points, reduced to Admiralty Chart Datum using tide gauge measurements, and interpolated onto a 20-m-cell-size digital terrain model. The data set has a horizontal accuracy of ±20 m and vertical accuracy of ±10 cm. East–west and east northeast–west southwest striping observed in bathymetry images is an artefact of survey vessel tracks. See Methods for additional details.

Palaeohydrological estimation of flood discharge. We estimated the peak discharge of the Northern Palaeovalley using the uniform flow Manning's equation[31], with the slope, channel width and channel depth derived from the sonar bathymetry (see Methods for details of assumptions and uncertainties in the calculation). By substituting mean flow velocity from Manning's equation in the continuity equation, flow discharge at a discrete point was calculated using $Q = dWn^{-1}R^{2/3}S^{1/2}$, where Q is the discharge ($m^3 s^{-1}$), d is the mean channel depth, W is the mean channel width, n is Manning's roughness coefficient, R is the hydraulic radius (taken to be mean channel depth) and S is the down-flow slope of the channel bed ($m m^{-1}$). We used a Manning's n of 0.04 estimated for rock-bound channels[32], consistent with previous studies of floods in bedrock channels[13,14]. We estimated discharges for the two episodes of flood erosion identified and assumed bankfull flow in each case (that is, water entirely filled each palaeochannel to its rims): event one involved erosion from the trim line to the upper surface of the bench, and event two eroded to the inner channel floor. For event one (at line of cross-section A–A′, Fig. 2a), $d = 20$ m, $W = 20{,}000$ m (a minimum width because the southern channel margin is not observed in the survey area, and is likely to be much greater) and $S = 0.0002$. For event two, we estimated discharges for a maximum and minimum width of the valley. A maximum width of $W = 12{,}000$ m was measured immediately downstream of island M1, and a minimum width of $W = 7{,}000$ m was obtained at the prominent constriction in the valley just upstream of M1. Channel depth was measured as $d = 15$ m and slope was $S = 0.0002$. From these measurements, we estimate peak discharge Q_{event1} of $\sim 1 \times 10^6 m^3 s^{-1}$, and Q_{event2} between ~0.2 and $0.4 \times 10^6 m^3 s^{-1}$.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 September 2006; accepted 11 June 2007.

1. Baker, V. R. The study of superfloods. *Science* **295,** 2379–2380 (2002).
2. Smith, A. J. A catastrophic origin for the paleovalley system of the eastern English-Channel. *Mar. Geol.* **64,** 65–75 (1985).
3. Roep, T. B., Holst, H., Vissers, R. L. M., Pagnier, H. & Postma, D. Deposits of southward-flowing, Pleistocene rivers in the Channel region, near Wissant, N.W. France. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **17,** 289–308 (1975).
4. Gibbard, P. L. in *Island Britain: a Quaternary Perspective* (ed. Preece, R. C.) 15–26 (Geological Society Special Publication, London, 1995).
5. Destombes, J. P., Shephardthorn, E. R., Redding, J. H. & Morzadeckerfourn, M. T. Buried valley system in strait of Dover. *Phil. Trans. R. Soc. Lond. A* **279,** 243–256 (1975).
6. Kellaway, G. A., Redding, J. H., Shephardthorn, E. R. & Destombes, J. P. The quaternary history of the English Channel. *Phil. Trans. R. Soc. Lond. A* **279,** 189–218 (1975).
7. Preece, R. C. (ed.) *Island Britain: a Quaternary Perspective* (Geological Society Special Publication, London, 1995).
8. Hamilton, D. & Smith, A. J. The origin and sedimentary history of the Hurd Deep, English Channel, with additional notes on other deeps in the western English Channel. *Mem. Bur. Rech. Geol. Min.* **79,** 59–78 (1972).
9. Dingwall, R. G. Sub-bottom infilled channels in an area of eastern English-Channel. *Phil. Trans. R. Soc. Lond. A* **279,** 233–241 (1975).
10. Auffret, J. P., Alduc, D., Larsonneur, C. & Smith, A. J. Maps of the paleovalleys and of the thickness of superficial sediments in the eastern English-Channel. *Ann. Inst. Oceanogr.* **56,** 21–35 (1980).
11. Hamblin, R. J. O. *et al. United Kingdom Offshore Regional Report: the Geology of the English Channel* (HMSO for the British Geological Survey, London, 1992).
12. British. Geological Survey *Wight. 1:250,000 (Solid Geology) map* (British Geological Surveys, Edinburgh, 1995).
13. Baker, V. R. & Nummedal, D. (eds) *The Channeled Scabland; a Guide to the Geomorphology of the Columbia Basin, Washington* (National Aeronautics and Space Administration, Washington DC, 1978).
14. Baker, V. R. *The Channels of Mars* (Univ. Texas Press, Austin, Texas, 1982).
15. Komar, P. D. Shapes of streamlined islands on Earth and Mars — experiments and analyses of the minimum-drag form. *Geology* **11,** 651–654 (1983).
16. Gibbard, P. L. The history of the great northwest European rivers during the past 3 million years. *Phil. Trans. R. Soc. Lond. B* **318,** 559–602 (1988).
17. Ehlers, J. & Gibbard, P. L. (eds) *Quaternary Glaciations—Extent and Chronology. Part I: Europe* 251–270 (Elsevier, Amsterdam, 2004).
18. Meijer, T. & Preece, R. C. in *Island Britain: a Quaternary Perspective* (ed. Preece, R. C.) 89–110 (Geological Society Special Publication, London, 1995).
19. Keen, D. H. in *Island Britain: a Quaternary Perspective* (ed. Preece, R. C.) 63–74 (Geological Society Special Publication, London, 1995).
20. O'Connor, J. E., Grant, G. E. & Costa, J. E. in *Ancient Floods, Modern Hazards—Principles and Applications of Paleoflood Hydrology* (ed. House, P.) 359–385 (American Geophysical Union, Washington DC, 2002).
21. Antoine, P. *et al.* The Pleistocene rivers of the English channel region. *J. Quat. Sci.* **18,** 227–243 (2003).
22. Lericolais, G., Auffret, J. P. & Bourillet, J. F. The Quaternary Channel River: seismic stratigraphy of its palaeo-valleys and deeps. *J. Quat. Sci.* **18,** 245–260 (2003).
23. Menot, G. *et al.* Early reactivation of European rivers during the last deglaciation. *Science* **313,** 1623–1625 (2006).
24. White, M. J. & Shreve, D. C. Island Britain–peninsula Britain: palaeogeography, colonisation and the lower palaeolithic settlement of the British Isles. *Proc. Prehist. Soc.* **66,** 1–28 (2000).
25. Parfitt, S. A. *et al.* The earliest record of human activity in northern Europe. *Nature* **438,** 1008–1012 (2005).
26. Ashton, N. & Lewis, S. Deserted Britain: declining populations in the British Late Middle Pleistocene. *Antiquity* **76,** 388–396 (2002).
27. White, M., Scott, B. & Ashton, N. The Early Middle Palaeolithic in Britain: archaeology, settlement history and human behaviour. *J. Quat. Sci.* **21,** 525–541 (2006).
28. Sutcliffe, A. J. in *Island Britain: a Quaternary Perspective* (ed. Preece, R. C.) 127–140 (Geological Society Special Publication, London, 1995).
29. Barber, D. C. *et al.* Forcing of the cold event of 8,200 years ago by catastrophic drainage of Laurentide lakes. *Nature* **400,** 344–348 (1999).
30. Broecker, W. Was the Younger Dryas triggered by a flood? *Science* **312,** 1146–1148 (2006).
31. Manning, R. On the flow of water in open channels and pipes. *Trans. Inst. Civ. Eng. Ir.* **20,** 161–207 (1891).
32. Chow, V. T. *Open Channel Hydraulics* (McGraw-Hill, New York, 1959).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The bathymetry surveys were funded by the Maritime and Coastguard Agency under the UK Civil Hydrography Programme; we thank J. Collins and R. Spillard for support. Data acquisition for Fig. 4b was funded by the Aggregates Levy Sustainability Fund through English Heritage and by the Joint Research Equipment Initiative (HEFCE/HEFCW). We thank B. Coakley, A. Densmore, R. S. Anderson, P. A. Allen, C. Paola, S. Parfitt, R. Preece and N. Ashton for discussions, and V. Baker and P. Gibbard for their comments.

Author Contributions S.G. and J.S.C. analysed the bathymetry data and wrote the paper. G.P. compiled and processed the data, together with A.P.-F. who also aided the analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.G. (s.gupta@imperial.ac.uk).

METHODS

Multibeam data acquisition and processing. High-resolution swath bathymetric data were collected in March 2003 across an 8×17 km patch of sea floor off Littlehampton, UK, using a Reson 8101 multibeam echosounder, and were used to generate Fig. 4b. Operating at 240 kHz, the system formed 110 $1.5^\circ \times 1.5^\circ$-beams over a 150° swath, producing a footprint of 0.5–3.5 m^2 in water depths ranging from 18 m to 55 m. The Reson 8101 was mobilized on the 12-m-long catamaran *Explorer of Portsmouth* using an over-the-side pole mounting. Vehicle motion was measured using an Applanix POS MV 220 integrated motion sensor, gyrocompass and GPS positioning system. A CSI Wireless differential GPS receiver using Trinity House RTCM corrections determined vessel position with sub-metre accuracy. The Reson SVP-C sensor continuously monitored sound speed at the transducer head, and a Reson SVP-14 sound-speed profiler was lowered to measure velocity profiles to the seabed.

Offsets between all of the sensors were carefully measured, and standard hydrographic patch test and calibration procedures were followed to correct for any misalignment between the sonar head and the survey vessel. A Valeport pressure-type tide gauge was installed in Brighton Marina, and measurements were made at ten-minute intervals throughout the survey period. A Reson 6042 data acquisition system was used to digitally acquire data, integrate data from the ancillary sensors, and to store raw data files. Data from the 6042 were exported into an eXtended Triton Format for processing with CARIS HIPS, which produced cleaned and gridded data. The data were gridded at a 3-m cell size.

Data visualization. The coloured and shaded relief bathymetry and onshore digital terrain images used in Figs 1–4 were generated using Interactive Visualisation Systems' (IVS 3D) visualization software Fledermaus.

Uncertainties and assumptions in estimation of flood discharge. There are many uncertainties inherent in reconstructing palaeo-flood discharges from the geological record, so our discharge estimates can be considered only a first approximation. Although more sophisticated methods for modelling flow conditions are available, these require additional assumptions and thus are not warranted for a preliminary understanding of flood discharge. In the future, more detailed characterization of the topography of the valleys (for example, more accurate determination of flood-stage indicators) and identification of sedimentary bedforms and deposits associated with flooding may permit more accurate reconstruction of flood discharges.

In our discharge estimation, we applied Manning's equation at discrete points along the channel (for example, cross-section A–A′, Fig. 2b) using the conveyance-slope method. Estimation of discharge and palaeoflow velocity requires estimates of channel cross-section dimensions, energy slope (S) and Manning's roughness coefficient (n). Below, we discuss some of the key assumptions and uncertainties in our discharge reconstruction.

It is uncertain whether the empirical Manning equation is appropriate for flows that are several orders of magnitude larger than any flow that has been directly measured. For example, the Manning equation assumes uniform flow, which is highly unlikely during catastrophic floods. Reference 33 indicated that there is no way to evaluate such assumptions; however, they noted that the Manning equation has proved appropriate for flows over several orders of magnitude (albeit much smaller). Our estimates thus enable an order-of-magnitude approximation of palaeo-flood discharge, but should nevertheless be treated as preliminary.

Key assumptions in discharge estimation. Accurate definition of the geometry and dimensions of channels is important for palaeo-discharge estimation. The channel dimensions were obtained from topographic profiles determined from the bathymetric data. The two-level channel geometry of the Northern Palaeovalley is interpreted as resulting from two episodes of valley incision caused by separate flooding events. We thus treat the valley as being comprised of two channels, with the second having incised into the base of the first, and we estimate discharges for each of the flood events. Event one carved the valley from the upper edge of the northern valley margin (the trim line TL) to the level of the bedrock bench (Fig. 2b). Clear evidence for high water marks is not apparent in our data; however, multibeam data from further to the northeast of the valley (Fig. 4b and Supplementary Fig. 1) indicate sharp erosion at the upper edge of the valley that we interpret as evidence of flood erosion. There is no evidence that the flood overspilled the rims of the valley, so the valley edge provides an upper flood stage indicator. Thus, we consider flow to have been bankfull between the bench level and the valley edge. For event one, we cannot constrain the width of the valley accurately because only one valley margin is observed in the survey area. We use a width of 20 km along the line of section A–A′; however, the valley width was almost certainly much greater (Fig. 2a).

Event two eroded the valley from the level of the bedrock bench down to the present floor of the channel. The cross-sectional morphology of this channel and the truncation of drainage nets on island M1 suggests that, for event two, flow was bankfull during flood erosion. That the upper edge of the Northern Palaeovalley and the bedrock bench maintain similar elevations, and that we do not see evidence of additional terracing, suggest that our assumption of bankfull flow may be appropriate. The geometry of the channels at each of these stages is assumed to represent the palaeogeometry of the channel at the time of flooding. The first flood event may initially have occupied a 'normal' fluvial bedrock channel, and thus the depth of flood erosion of the first channel may be lower than indicated. However, discharge estimates are more affected by the great width of the flows. It is very unlikely that a pre-flood river attained such a great width. Another assumption is that the measured cross-section approximates channel geometry at peak flood stages, that is, there has not been any significant post-peak downcutting, filling or valley widening. The geomorphic evidence indicates that this is the case.

Application of Manning's equation requires determination of the channel's hydraulic radius R. In channels that are much wider than they are deep, the depth of the channel can be used as a proxy for R. Application of Manning's equation also requires determination of the energy slope of the flood flow. The water surface slope and the resulting energy slope could not readily be determined so we used the present channel gradient as a proxy. This slope is probably less than the energy slope, and thus the discharge is likely to be an underestimate. We estimated channel slope by using the average gradient of the bedrock bench.

The selection of a Manning roughness coefficient (n) for the English Channel valleys is difficult because the empirical Manning equation derived for normal rivers has to be scaled to the large depths of flood flows. We used a value of $n = 0.040$ from the empirical table for rock-bound channels described in ref. 32. References 13 and 34 used this value of n to estimate roughness for basalt bedrock in the Channeled Scabland. The Chalk bedrock in the Northern Palaeovalley may have a lower value of n, but this would only serve to increase the discharge values obtained.

The application of both Chézy's equation and the Darcy–Weisbach equation to estimate palaeo-flow velocities and subsequently flood discharges in the Northern Palaeovalley yields results of a similar order of magnitude to those obtained using the Manning equation. We are thus confident in the approximate range of the discharges estimated.

33. O'Conner, J. E. & Baker, V. R. Magnitudes and implications of peak discharges from glacial Lake Missoula. *Geol. Soc. Am. Bull.* **104**, 267–279 (1992).
34. Baker, V. R. *Paleohydrology and Sedimentology of Lake Missoula Flooding in Eastern Washington* (Geological Society of America, Boulder, Colorado, 1973).

LETTERS

The effect of ancient population bottlenecks on human phenotypic variation

Andrea Manica[1], William Amos[1], François Balloux[2] & Tsunehiko Hanihara[3]

The origin and patterns of dispersal of anatomically modern humans are the focus of considerable debate[1–3]. Global genetic analyses have argued for one single origin, placed somewhere in Africa[4–7]. This scenario implies a rapid expansion, with a series of bottlenecks of small amplitude, which would have led to the observed smooth loss of genetic diversity with increasing distance from Africa. Analyses of cranial data, on the other hand, have given mixed results[8–12], and have been argued to support multiple origins of modern humans[2,9,12]. Using a large data set of skull measurements and an analytical framework equivalent to that used for genetic data, we show that the loss in genetic diversity has been mirrored by a loss in phenotypic variability. We find evidence for an African origin, placed somewhere in the central/southern part of the continent, which harbours the highest intra-population diversity in phenotypic measurements. We failed to find evidence for a second origin, and we confirm these results on a large genetic data set. Distance from Africa accounts for an average 19–25% of heritable variation in craniometric measurements—a remarkably strong effect for phenotypic measurements known to be under selection.

The origin of anatomically modern humans has been the focus of much heated debate[1,3]. Recent large scale genetic analyses[4–7] seem to support the idea that all modern humans originated from a single location (the 'single origin' hypothesis). More specifically, all studies point to Africa as the putative cradle of modern humans. If rapid, the expansion out of Africa would imply progressive loss of genetic diversity through a series of founder events (bottlenecks), a prediction that has recently received empirical support[4–6]. Heterozygosity declines monotonically with distance from east Africa, with South American populations carrying 64% of the neutral variability (as measured from microsatellites) found in African populations. This view is further supported by some archaeological and anthropological evidence[1]. However, studies of craniometric data have yielded mixed results[8–12], and the presence of archaic human-like traits in skulls that would be otherwise classified as *Homo sapiens* in several continents has been interpreted as evidence for multiple origins (the 'multiregional' hypothesis)[2,12].

An important step towards an unequivocal answer regarding the number of origins of modern humans would be to analyse the phenotypic (cranial) measurements using the same approach used for genetic traits. The alternative models (single origin and multiregional) make clear predictions about how craniometric diversity should be distributed. Under the single origin model, we expect to find a monotonic decrease in phenotypic variability analogous to that seen for genetic traits (unless the sampling process is so strong as to destabilize canalization through the loss of genetic diversity)[13,14]. In contrast, multiple origins should lead to several clines, the magnitude of each cline being determined by the relative contribution of its origin.

To test these predictions, we used an exceptional data set of 4,666 male skulls measured for 37 morphometric characteristics (Supplementary Table S1) and drawn from 105 populations (Fig. 1; Supplementary Table S2)[15,16]. A minimum sample size of 15 individuals was enforced (median size, 36) and skulls older than 2,000 years were excluded to avoid any bias in the quality of the material. It is well known that some skull measurements correlate with climate, implying natural selection[17,18]. Consequently, before considering

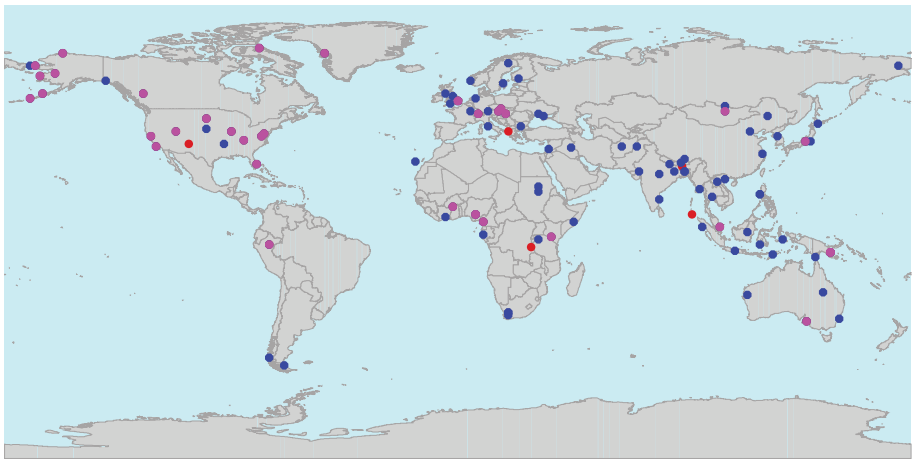


Figure 1 | Origins of samples. Map of locations of populations from which male (blue) and female (red) skulls were collected. Locations from which skulls of both sexes were collected are marked in purple.

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. [2]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. [3]Department of Anatomy, Saga Medical School, Saga 849-8501, Japan.

the effect of ancient demography, we investigated the effect of three key climatic variables on the within-population cranial variability: maximum and minimum temperature, and average annual precipitation. For each population, we estimated the mean standardized phenotypic variance[19] and fitted it in a linear model with climate variables (maximum temperature, minimum temperature and precipitation, with all possible interactions) as predictors (see Methods). Backwards stepwise elimination selected the interaction between maximum temperature and precipitation (ΔBayesian Information Criterion, ΔBIC = −2.2) as the best predictor of cranial variability. The minimal model including this interaction was used as the starting point for the investigation of any effect of ancient demography, effectively assuming that climate was the most parsimonious explanation for any global pattern, and that ancient demography had to explain variance that had not already been accounted for by climate. This is arguably a very conservative approach, given that a strong pattern of isolation by distance in craniometric inter-population differences has been found[18], and that these authors argued that selection through climate is insufficient to erase ancient demographic signals.

We next searched the globe for the putative origin giving the strongest relationship between within-population phenotypic variability (corrected for climate) and distance on land[5,20] (see Methods). This strongest cline originates in central/southern Africa, and could be either a single origin or the main origin in a multiregional scenario (Figs 2a and 3; effect of distance from the centroid of the likely origins after correcting for climate: ΔBIC = −12.5; R^2 of plot against distance, 14.0). A similar analysis of the 789 autosomal neutral microsatellites from the 54 populations of the HGDP-CEPH panel[21,22], using heterozygosity as a measure of genetic variability, gives almost identical results (Fig. 2b; showing patterns similar to those obtained by Ramachandran *et al.*[6] and Ray *et al.*[7], who used previous versions of the HGDP-CEPH data set and somewhat different analytical approaches). The only noticeable difference between the estimates based on phenotypic and genetic traits is that the latter does not include south Africa amongst the most likely origins.

To test the multiregional hypothesis, we sought a second, non-African origin capable of increasing the explanatory power of the model (see Methods). Adding distances from other non-African origins did not improve our models (for either phenotypic or genetic traits), supporting Ray *et al.*[7], who used a spatially explicit stochastic population model and also failed to find evidence for the multiregional hypothesis. Our approach therefore suggests that a multiple origin is unlikely. However, we cannot distinguish between single and multiple exoduses from Africa, because both scenarios would lead to a major cline from Africa. Depending on the exact timing and extent of the multiple exoduses, we could expect several subtly different patterns, and neither the genetic nor phenotypic data sets are currently large enough to investigate this level of detail. Also, very localized episodes of admixture between anatomically modern and archaic humans might go undetected if they left no signature in present day modern humans[2].

How strong are the patterns for individual cranial measurements? As we could not identify a precise location for the African origin of humans, we chose a centroid of the likely craniometric origins, and the possible influence of outliers was minimized by using robust regression[23] (see Methods). We considered both a model with distance as the only predictor, and one with a correction for climate. Out of 37 cranial measurements, 34 showed a decline in variability with distance from Africa (binomial test, $P < 0.001$; Supplementary Table S3), and 12 of these (18 if we did not correct for climate) were significant after correction for multiple testing with a false discovery rate procedure[24]. Although mean R^2 was 6.3% (7.0% if we did not correct for climate), some measurements showed a strong demographic signal (Fig. 4; Supplementary Table S3). A further test of how reliable these results are is to repeat the analysis on an equivalent but smaller data set of female skulls, composed of 1,579 individuals from 39 populations (Supplementary Table S2). The results are very similar to those found in males (Supplementary Table S4), with a mean R^2 of 9.1% for distance (8.9% without correcting for climate). Furthermore, slopes of measurement coefficient of variation over distance for females are almost identical to those obtained for males (best regression line: female slope for measurement $x = 1.00 \times$ male slope for measurement x, $R^2 = 0.77$, $F_{1,36} = 124.9$, $P < 0.001$).

Given that distance from Africa explains over 87% of variance in heterozygosity at neutral microsatellite markers[5], the equivalent values for morphometric measurements might seem disappointing. However, phenotypes are only partially determined by genotypes, as the environment is also playing an important role. Heritability, h^2, defines the fraction of variance in a trait affected by genetics, and this represents an upper ceiling for the size of the demographic signal that can be detected. Estimates of heritability for cranial measurements are rare,

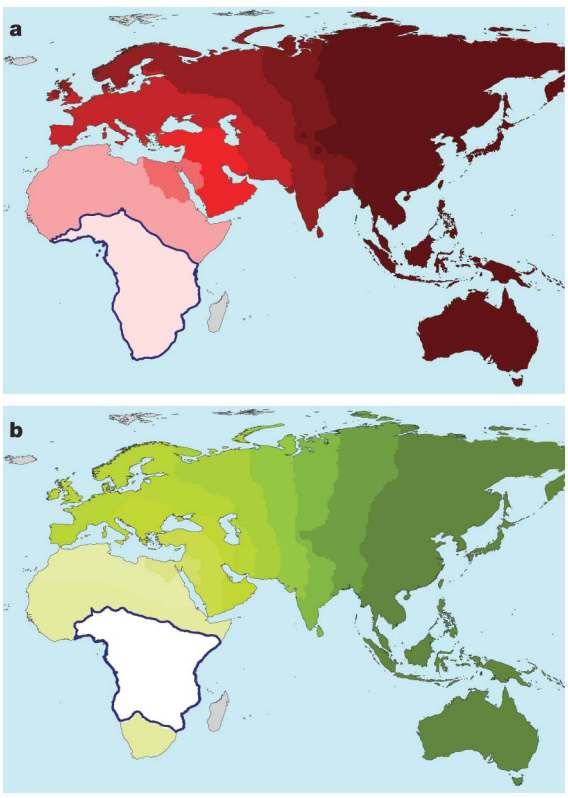


Figure 2 | Likely origin of anatomically modern humans. a, b, Maps showing the likely location of a single origin for phenotypic (**a**) and genetic (**b**) data. Lighter colours represent better fits of models of variability as predicted by distance from a location. The area in each panel containing the most likely origin is enclosed by a blue line. Areas of the world not investigated as possible origins (such as Iceland and Madagascar, which would require substantial land bridges to the main continents) are shown in grey.

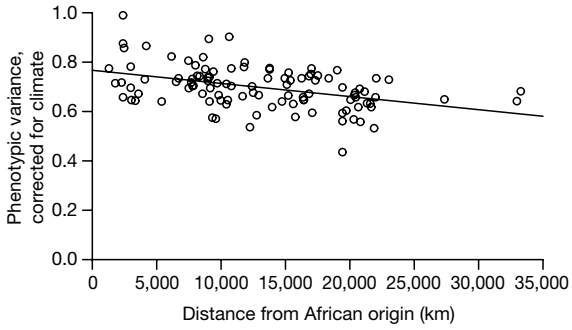


Figure 3 | Phenotypic variability versus distance from Africa. Relationship between mean phenotypic variability (corrected for climate) for male skulls and distance from the putative African origin (represented by the centroid of likely origins). The solid line represents the best fit from the regression.

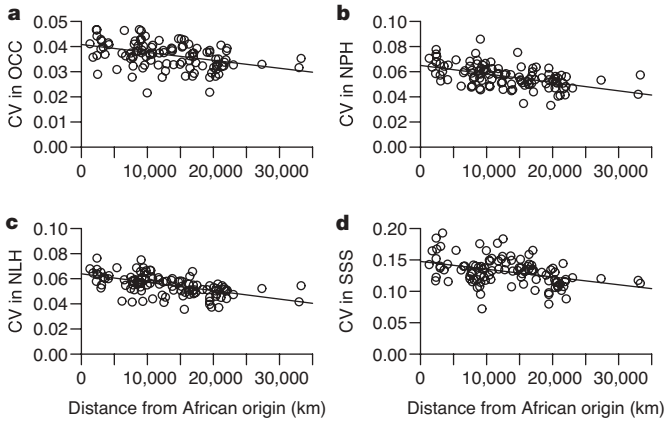


Figure 4 | Phenotypic variability of individual traits. Relationships between phenotypic variability for four measurements (expressed as coefficient of variation, CV, corrected for climate) from male skulls and distance from the putative African origin (represented by the centroid of locations of origins). The measurements are: **a**, lambda-opisthion chord (OCC); **b**, nasion prosthion height (NPH); **c**, nasal height (NLH); and **d**, zygomaxillary subtense (SSS).

but values for 19 of the measurements included in our analysis have been recently derived[25], and these correlate significantly with the proportion of variance explained by distance from Africa (males $\rho = 0.56$, $P = 0.012$). On average, $R^2/h^2 = 0.20$ (s.e. = 0.06) for males and $R^2/h^2 = 0.20$ (s.e. = 0.06) for females (0.24 ± 0.06 and 0.25 ± 0.06, respectively, if we did not correct for climate), indicating that the proportion of explained variance is at least a fifth of the heritability. Interestingly, this figure is similar to the 17–35% of variance explained by distance from Africa for class I MHC genes which are under selection by infectious diseases[26], just as climatic (and potentially other) factors select for skull shape. Thus, after allowing for the impact of heritability and selection, the signal of ancient demography in human skull variability should properly be seen as remarkably strong.

METHODS SUMMARY

Thirty-seven morphometric cranial measurements (Supplementary Table S1) were made on 4,666 male skulls drawn from 105 populations (Fig. 1; Supplementary Table S2)[15,16]. After correcting for climate, we looked for the likely origin of human diversity by investigating the relationship between phenotypic diversity (scored as mean standardized phenotypic variance[19]) and distances on land from locations across the globe. We then tested for the presence of a second origin by testing whether adding distances to putative locations improved the single origin model. We also assessed the effect of ancient demography on variability in individual measurements for both male skulls and an additional data set of 1,579 female skulls from 39 populations, and compared patterns found in the two sexes. Finally, we investigated the magnitude of the ancient demographic signal in the context of the heritability of individual measurements, as heritability provides an upper limit to any signature left by genetic factors.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 January; accepted 22 May 2007.

1. Mellars, P. Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313,** 796–800 (2006).
2. Trinkaus, E. Early modern humans. *Annu. Rev. Anthropol.* **34,** 207–230 (2005).
3. Mellars, P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc. Natl Acad. Sci. USA* **103,** 9381–9386 (2006).
4. Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79,** 230–237 (2006).
5. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15,** R159–R160 (2005).
6. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102,** 15942–15947 (2005).
7. Ray, N., Currat, M., Berthier, P. & Excoffier, L. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res.* **15,** 1161–1167 (2005).
8. Grine, F. E. *et al.* Late Pleistocene human skull from Hofmeyr, South Africa, and modern human origins. *Science* **315,** 226–229 (2007).
9. Wolpoff, M. H., Hawks, J. & Caspari, R. Multiregional, not multiple origins. *Am. J. Phys. Anthropol.* **112,** 129–136 (2000).
10. Brauer, G., Collard, M. & Stringer, C. On the reliability of recent tests of the Out of Africa hypothesis for modern human origins. *Anat. Rec. A* **279,** 701–707 (2004).
11. Lahr, M. M. The multiregional model of modern human origins — A reassessment of its morphological basis. *J. Hum. Evol.* **26,** 23–56 (1994).
12. Wolpoff, M. H. in *The Human Revolution: Biological Perspectives in the Origins of Modern Humans* (eds Mellars, P. & Stringer, C.) 62–108 (Princeton Univ. Press, Princeton, 1989).
13. Fowler, K. & Whitlock, M. C. The distribution of phenotypic variance with inbreeding. *Evolution Int. J. Org. Evolution* **53,** 1143–1156 (1999).
14. Frankham, R. Do island populations have less genetic variation than mainland populations? *Heredity* **78,** 311–327 (1997).
15. Hanihara, T. & Ishida, H. Metric dental variation of major human populations. *Am. J. Phys. Anthropol.* **128,** 287–298 (2005).
16. Hanihara, T. & Ishida, H. Os incae: variation in frequency in major human population groups. *J. Anat.* **198,** 137–152 (2001).
17. Roseman, C. C. Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. *Proc. Natl Acad. Sci. USA* **101,** 12824–12829 (2004).
18. Relethford, J. H. Boas and beyond: Migration and craniometric variation. *Am. J. Hum. Biol.* **16,** 379–386 (2004).
19. Relethford, J. H. & Blangero, J. Detection of differential gene flow from patterns of quantitative variation. *Hum. Biol.* **62,** 5–25 (1990).
20. Manica, A., Prugnolle, F. & Balloux, F. Geography is a better determinant of human genetic differentiation than ethnicity. *Hum. Genet.* **118,** 366–371 (2005).
21. Rosenberg, N. A. Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* **70,** 841–847 (2006).
22. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–2385 (2002).
23. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, New York, 2002).
24. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
25. Carson, E. A. Maximum likelihood estimation of human craniometric heritabilities. *Am. J. Phys. Anthropol.* **131,** 169–180 (2006).
26. Prugnolle, F. *et al.* Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* **15,** 1022–1027 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge J. Goudet for discussions. The Biotechnology and Biological Sciences Research Council provided financial support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M. (am315@cam.ac.uk).

METHODS

Data sets. Thirty-seven morphometric cranial measurements (Supplementary Table S1) were measured in 6,245 skulls (4,666 male and 1,579 female), drawn from 105 and 39 populations for males and females, respectively (Fig. 1; Supplementary Table S2). Details of the measurements are provided elsewhere[15,16]. Mid-oceanic populations were excluded because their origins are often unclear. Also excluded were samples over 2,000 years old, as deterioration might lead to biased estimates of phenotypic variability. Only the male data set was large enough to discriminate among potential origins of diversity. Consequently, the female data set was used for comparison when estimating the strength of the clines detected using male skulls. Genetic data come from the latest version of the HGDP-CEPH panel[21,22], which includes 971 individuals belonging to 54 populations and typed for 789 neutral autosomal microsatellites. Climate data (minimum and maximum temperature, and average precipitation) were obtained from WORDCLIM[27], as sets of global climatic GIS layers with a 30 arcsec resolution.

Statistical analysis. Total within-population phenotypic variability was computed as the mean standardized phenotypic variance over all measurements. Following ref. 19, this measure can be computed as the trace of the variance/covariance matrix of standardized values divided by the number of traits. Variability of individual measurements within each population was computed as the coefficient of variation (CV, the ratio of the standard deviation to the mean). This dimensionless measure allows comparisons between different measurements. Realistic geographic distances between locations were computed as the shortest route through landmasses that avoid areas with a mean altitude over 2,000 m and assuming the following land bridges: a single connection between Africa and Eurasia via a route through the Sinai to the Levant, the Bering Strait between Eurasia and the Americas, and connections between the Malaysian Peninsula to Melanesia and Oceania[5,20].

To model the effect of climate on craniometric measurements (male skulls only), we fitted a linear model with mean standardized phenotypic variance as the response and three climatic variables and all their possible interactions as predictors. Starting from this full model, we then found a minimal model for climate by backwards stepwise elimination using the BIC[28].

The presence of a primary cline in diversity was tested by fitting an additional predictor to the minimal climate model, namely distance from a range of potential origins distributed at intervals of 5 degrees latitude/longitude across the whole of Africa, Eurasia and Australia. As we effectively had to compare hundreds of similar models with only a slight difference in the values of the predictor variable (distances from different locations), we used BIC to select origins that gave similarly good fits. Models within four units of the best model have ''considerable support''[29], thus providing a suitable envelope for the likely real origin. Having identified an African origin, we searched for a second origin by fitting distances from origins outside Africa as an additional predictor to the best African model. This approach was then repeated using the genetic data set, with heterozygosity as the response variable in a linear model.

The strength of the relationship between variability of individual measurements and distance from Africa was explored by considering individual measurements in both males and females. Distances from a centroid of likely origins were fitted on top of the minimal climate model with the CV of measurements as individual responses. Visual inspection of the relationship between CV and geographic distance revealed outliers in several measurements. To account for heteroskedasticity without resorting to *ad hoc* removal of populations, we used robust regression with MM-estimators[23] fitted in R[30] using the robustbase package. This technique allows the estimation of linear models that are not affected by the outliers. We fitted models both with and without correcting for climate. The proportion of variance explained by distance from Africa when correcting for climate was estimated by subtracting the R^2 of the model including distance and climate from the R^2 of the climate model alone. Owing to the large number of nested tests, it is difficult to formulate exact P values. To provide a measure of goodness of fit, we computed local false discovery rates using the smoothing spline approach in ref. 24 to estimate η_0, the proportion of null P values (computations used the fdrtools package in R[30]).

27. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).
28. Crawley, M. J. *Statistical Computing: An Introduction to Data Analysis Using S-Plus* (Wiley, London, 2002).
29. Burnham, K. P. & Anderson, D. R. *Model Selection and Inferences* (Springer, New York, 1998).
30. R Development Core Team (R Foundation for Statistical Computing, Vienna, 2006).

LETTERS

Positive darwinian selection at the imprinted *MEDEA* locus in plants

Charles Spillane[1,2], Karl J. Schmid[3]†, Sylvia Laoueillé-Duprat[2], Stéphane Pien[1], Juan-Miguel Escobar-Restrepo[1], Célia Baroux[1], Valeria Gagliardini[1], Damian R. Page[1], Kenneth H. Wolfe[4] & Ueli Grossniklaus[1]

In mammals and seed plants, a subset of genes is regulated by genomic imprinting where an allele's activity depends on its parental origin. The parental conflict theory suggests that genomic imprinting evolved after the emergence of an embryo-nourishing tissue (placenta and endosperm), resulting in an intragenomic parental conflict over the allocation of nutrients from mother to offspring[1,2]. It was predicted that imprinted genes, which arose through antagonistic co-evolution driven by a parental conflict, should be subject to positive darwinian selection[3]. Here we show that the imprinted plant gene *MEDEA* (*MEA*)[4,5], which is essential for seed development, originated during a whole-genome duplication 35 to 85 million years ago. After duplication, *MEA* underwent positive darwinian selection consistent with neo-functionalization and the parental conflict theory. *MEA* continues to evolve rapidly in the out-crossing species *Arabidopsis lyrata* but not in the self-fertilizing species *Arabidopsis thaliana*, where parental conflicts are reduced. The paralogue of *MEA*, *SWINGER* (*SWN*; also called *EZA1*)[6], is not imprinted and evolved under strong purifying selection because it probably retained the ancestral function of the common precursor gene. The evolution of *MEA* suggests a late origin of genomic imprinting within the Brassicaceae, whereas imprinting is thought to have originated early within the mammalian lineage[7].

Disruption of the imprinted *Arabidopsis MEA* gene, which encodes an Enhancer of zeste [E(z)]-related protein, leads to delayed development and over-proliferation of embryo and endosperm[4,8]. Together with *SWN* and *CURLY LEAF* (*CLF*), *MEA* forms a family of *E(z)*-like genes in *Arabidopsis*[9,10]. To gain insights into their evolutionary relationship, we investigated whether the *Arabidopsis E(z)*-like genes arose via duplication of large genomic blocks[11]. *MEA* (At1g02580) is a recently derived paralogue of *SWN* (At4g02020) located on a block duplication spanning 39 paralogues on chromosome 1 and 41 paralogues on chromosome 4 (Fig. 1a). The block duplication on which the *MEA* and *SWN* paralogues reside arose ~35–85 million years (Myr) ago as a result of a whole-genome duplication within the Brassicaceae lineage[11–13]. In contrast, the *CLF* gene (At2g23380) is located in a genomic region that exhibits no co-linearity with the regions containing *SWN* and *MEA* in either *Arabidopsis* (Fig. 1a) or rice (data not shown).

To investigate further these duplications, we included all known plant *E(z)*-like genes in a phylogenetic analysis (Fig. 1b). The presence of *CLF*-like and *SWN*-like genes in both monocotyledons and dicotyledons indicates a duplication separating *CLF* and the common ancestor of *SWN* and *MEA* before the divergence of these taxa ~200 Myr ago. In agreement with the block duplication data, we found no direct orthologues of *MEA* (as opposed to co-orthologues of both *MEA* and *SWN*) in the available sequences (including expressed sequence tags) from any species other than *Arabidopsis thaliana*. However, we obtained orthologues of both *MEA* and *SWN* in *Arabidopsis lyrata* (Supplementary Fig. 1). All previous phylogenetic analyses of the plant *E(z)*-like genes suggested that *MEA* is a basal outgroup to both *CLF* and *SWN*[6,10,14–17]. In contrast, our data reveal an old duplication between the *CLF* and *MEA/SWN* lineages, followed by a more recent duplication that produced *MEA* and *SWN* (Fig. 1b).

To analyse functional diversification of *E(z)*-like genes in *Arabidopsis*, we studied their expression and function during reproductive development. *MEA* is expressed in the synergids, egg and central cells of the embryo sac before fertilization, and in the embryo and endosperm during seed development[5]. To determine whether *MEA*, *SWN* and *CLF* have overlapping expression patterns in the embryo sac and developing seed, we performed comparative *in situ* hybridization analyses using gene-specific probes (Fig. 2 and Supplementary Fig. 2). Although the three genes have largely overlapping expression patterns, there are important differences: all three transcripts were detected in the synergids and egg cell (Fig. 2a, c, e); however, the expression of *SWN* was strongly reduced and that of *CLF* undetectable in the central cell compared with *MEA*, which showed strong expression (Fig. 2a). This difference was maintained during early seed development when *MEA* was detected in free nuclear endosperm (Supplementary Fig. 2a) but *SWN* and *CLF* were not (Supplementary Fig 2e, i). After fertilization, transcripts of all three genes were detected in the globular embryo and micropylar endosperm, with only *MEA* transcripts detected in the suspensor (Fig. 2b, d, f). The expression of all three genes decreased at the heart stage (data not shown) and they were no longer detectable in embryos of the torpedo stage (Supplementary Fig. 2o, s, w). Thus, in comparison to its paralogues *SWN* and *CLF*, *MEA* has a differential expression domain in the central cell and free nuclear endosperm, and also in the suspensor, both tissues thought to be involved in nutrient transfer to the developing embryo.

Mutations in *MEA* and other members of the *FERTILIZATION-INDEPENDENT SEED* (*FIS*) class of genes show characteristic pre- and post-fertilization phenotypes: endosperm proliferation in the absence of fertilization (*fis* phenotype) and maternal effect seed abortion[4,18–21]. As *MEA* is a maternally expressed imprinted gene, maternal effect seed abortion in *Arabidopsis mea* mutants can occur in heterozygous seeds that have maternally inherited a mutant *mea* allele, yet harbour a wild-type paternal *MEA* allele that is not expressed. To test whether the paternal allele of the *MEA* orthologue in *A. lyrata* is also not expressed in seeds, we analysed the relative expression levels of maternal and paternal *MEA* alleles in seeds

[1]Institute of Plant Biology & Zürich-Basel Plant Science Center, University of Zürich, CH-8008 Zürich, Switzerland. [2]Genetics & Biotechnology Lab, Department of Biochemistry & Biosciences Institute, University College Cork, Cork, Ireland. [3]Department of Genetics and Evolution, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str. 8, D-07745 Jena, Germany. [4]Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland. †Present address: Leibniz-Institute of Plant Genetics and Crop Plant Research, D-06466 Gatersleben, Germany.

generated from crosses between *A. thaliana* and *A. lyrata* (as pollen parent). Similar to the imprinted *MEA* locus in *A. thaliana*, the *MEA* orthologue in *A. lyrata* is not expressed from the paternal allele in developing seeds (Supplementary Table 1). This result is consistent with our findings that a paternally inherited *A. lyrata MEA* allele cannot rescue the *mea* seed abortion phenotype (data not shown), and suggests that *MEA* is imprinted in both of the sister species, *A. thaliana* and *A. lyrata*.

Because *MEA* shows an overlapping expression pattern with *SWN* and *CLF*, we tested whether *swn* or *clf* mutants are impaired in either embryo sac or seed development. It was recently shown that *SWN* and *MEA* have a redundant function with respect to the pre-fertilization *fis* phenotype[22]. In contrast, our analysis of *swn* and *clf* mutants alone and in combination with *mea* showed neither an impairment of post-fertilization seed development nor an enhancement of the *mea* seed abortion phenotype, respectively (Supplementary Table 2). As *swn*;*clf* double mutants also produce normal seeds[6], these results indicate that neither *SWN* nor *CLF* has a role in seed development. Because neither *SWN* nor *CLF* are essential for post-fertilization seed development, we propose that the new post-fertilization role of *MEA* in seed development was acquired within the past ~35–85 Myr.

We further proposed that the protein sequence of MEA evolved rapidly after its origin by selection-driven substitutions of amino acids. In contrast, SWN would have retained the ancestral function and is expected to have evolved under purifying selection. We investigated the neo-functionalization hypothesis by testing whether the ratio $\omega = d_N/d_S$ of non-synonymous (d_N) to synonymous (d_S) divergence[23] is higher for the lineage ancestral to *MEA* than for *SWN* (Fig. 1c). The *E(z)*-like genes from poplar and petunia predate the *SWN–MEA* duplication and were taken as pro-orthologues, *sensu* ref. 24. A two-ratio branch model that estimates a single ω-ratio for the branches leading to the pro-orthologues and to *SWN* (reflecting functional conservation), and a second ω-ratio for *MEA* (allowing functional diversification), provided a significantly better fit to the data than a one-ratio model with a single ω-ratio for the whole phylogeny ($P < 0.0001$; Supplementary Table 3). Three-ratio and free-ratio models were not better than the two-ratio model ($P > 0.05$), suggesting that most variation in selective constraint occurred after the divergence of *SWN* and *MEA*.

Because *E(z)*-like proteins consist of a mosaic of conserved and divergent domains, we analysed which amino acid residues evolved under positive selection during functional diversification using branch-site models (Supplementary Table 4). This analysis revealed no evidence for positive selection in *SWN* ($P = 0.99$) but was highly significant for *MEA* ($P < 0.0001$). The ω-ratio of positively selected codons in the ancestral branch of *MEA* was estimated to be 1.68,

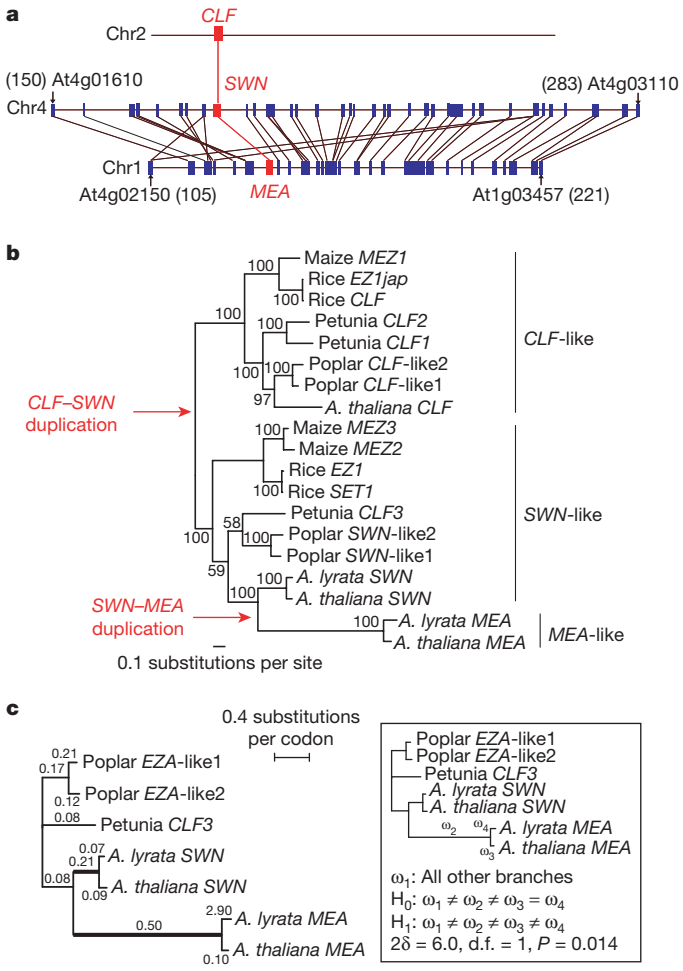


Figure 1 | *MEA* and *SWN* are paralogues. **a**, The imprinted *MEA* gene and its paralogue *SWN* lie on duplicated blocks of 0.457 megabases (Mb) and 0.690 Mb on chromosomes 1 and 4, respectively. The paralogon block spans 39 paralogous genes (blue blocks), including *MEA*, on chromosome 1, and 41 paralogues, including *SWN*, on chromosome 4. *CLF* is located on chromosome 2 in a genomic region exhibiting no co-linearity with the *MEA* and *SWN* paralogon blocks. **b**, Phylogenetic analysis of *E(z)*-like genes in higher plants. The tree was constructed with the protml package and rooted for the *CLF*—*SWN* duplication. **c**, Phylogenetic tree of *MEA*- and *SWN*-like genes from dicotyledonous species. The tree topology was obtained with protml, and the branch lengths (substitutions per codon) were calculated with codeml using Model M0. The numbers above branches indicate the ω-ratio and they were calculated with the free-ratio model. Note that the tree is unrooted.

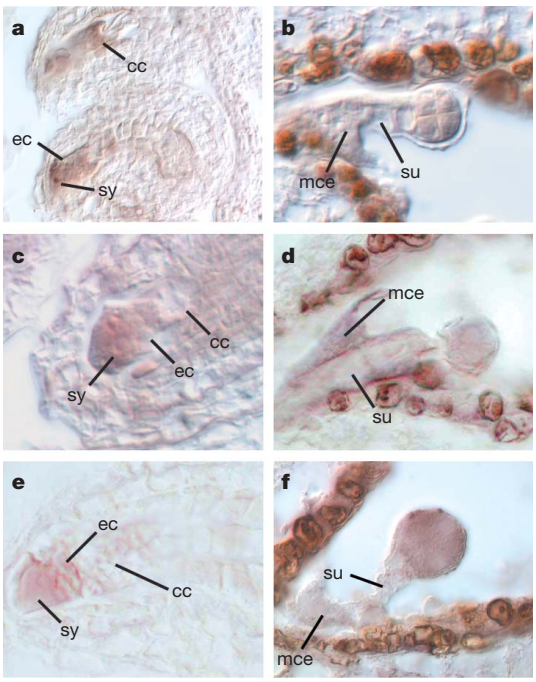


Figure 2 | Spatio-temporal expression patterns of *MEA*, *SWN* and *CLF* in the embryo sac and early seed assayed by *in situ* hybridization. **a**, **b**, Sections probed with anti-sense *MEA*. **c**, **d**, Sections probed with anti-sense *SWN*. **e**, **f**, Sections probed with anti-sense *CLF*. **a**, **c**, **e**, *MEA* (**a**), *SWN* (**c**) and *CLF* (**e**) transcripts accumulate in the synergids (sy) and in the egg cell (ec), whereas only *MEA* is expressed strongly in the central cell (cc). **b**, **d**, **f**, *MEA*, *SWN* and *CLF* transcripts were detected in the globular embryo and the micropylar endosperm (mce). **b**, At the globular stage, *MEA* transcripts are detected in the suspensor (su) in contrast to *SWN* and *CLF*. The strong staining seen in the endothelium is an artefact also observed in sense controls. For sense controls and a more detailed description of the expression patterns of *MEA*, *SWN* and *CLF* in the embryo sac and developing seed, see Supplementary Fig. 2.

which differs from a neutral model with ω fixed to 1.00 ($P = 0.026$). Therefore, the high ω-ratio does not result from relaxed constraints but from positive selection on *MEA*. The 74 codons with a posterior probability of positive selection >0.95 are located throughout the coding sequence (Supplementary Fig. 3), suggesting that positive selection was not restricted to any particular domain of the MEA protein. The numerous insertions and deletions of amino acids may also contribute to the functional divergence.

Within the genus *Arabidopsis*, *MEA*, but not *SWN*, continues to evolve under positive selection. The pairwise ω-ratio of *MEA* ($\omega = 0.75$) in *A. thaliana* and *A. lyrata* is higher than that of *SWN* ($\omega = 0.25$, $P < 0.0001$; Supplementary Table 5). A free-ratio model (Fig. 1c) indicates that *MEA* evolves under positive selection in the branch leading to *A. lyrata* ($\omega = 2.90$) but not in the *A. thaliana* branch ($\omega = 0.10$). A four-ratio model with independent ω-ratios for the *A. thaliana* and *A. lyrata MEA* genes provides a better fit to the data than a three-ratio model with a single ω-ratio for both branches ($2\delta = 6.0$, degrees of freedom = 1, $P = 0.014$; Fig. 1c), supporting positive selection in the *A. lyrata* branch. A test for differences in *SWN* between *A. thaliana* and *A. lyrata* was not significant (data not shown).

These results suggest that *MEA* is involved in a genomic conflict in *A. lyrata*, but not in the *A. thaliana* lineage where similar ω-ratios for *SWN* ($\omega = 0.09$) and *MEA* ($\omega = 0.10$) are observed (Fig. 1c). To test this hypothesis, we analysed intraspecific sequence variation at the *MEA* and *SWN* genes in *A. thaliana*, and at *MEA* in *A. lyrata* and its close, also out-crossing relative *A. halleri* (Supplementary Table 6). In the *A. thaliana* sample, total nucleotide diversity, π, is 1.5 times higher at *MEA* (0.0037) than at *SWN* (0.0022), and lower than the genome-wide average (0.007)[25]. The ratio of non-synonymous to synonymous polymorphisms, π_N/π_S, at *MEA* (0.259) is similar to *SWN* (0.183) and smaller than 1, indicating purifying selection. Several tests of neutral evolution failed to provide evidence for positive selection in *A. thaliana* (Supplementary Table 6); both loci appear to evolve neutrally under similar evolutionary constraints. In *A. lyrata* and *A. halleri*, similar patterns are observed for *MEA* (Supplementary Table 6). The π_N/π_S ratio is 1.25 and 0.58 in *A. lyrata* and *A. halleri*, respectively. Polymorphism levels are in the same range as observed for *A. thaliana*, but low in comparison to other *A. lyrata* genes located on the same chromosome. Tests of neutral evolution are not significant, although there is a slight excess of intermediate frequency polymorphisms in both species (Tajima's $D > 1$).

Positive darwinian selection on *MEA* occurring in the lineage leading to *A. lyrata* but not in *A. thaliana*, and a high π_N/π_S, is consistent with the parental conflict hypothesis for the evolution of imprinting. Within self-fertilizing *A. thaliana*, we expect differential selective pressures between maternal and paternal alleles of genes controlling resource allocation from mother to offspring (such as *MEA*) to be weaker. Because patterns of intraspecific sequence variation do not reject a neutral model in both the inbreeding and out-crossing species, selective pressures on *MEA* may be weak.

In mammals, imprinted gene clusters may have been linked together on one or a few ancestral chromosome(s), arguing for a common mechanistic origin of imprinting early in mammalian evolution[7]. In contrast, imprinting of *MEA* within the *E(z)*-like gene family arose late in the evolution of flowering plants, as *MEA*-like genes are restricted to the Brassicaceae. We propose that *MEA* became imprinted after it arose through a block duplication, possibly because of a need for dosage compensation after it acquired a new function. *SWN* and *CLF* have growth-regulating activity in the seedling, as demonstrated by aberrant growth of *swn;clf* double mutants after germination. Our studies on the evolution, function and expression of the *E(z)*-like genes suggest that *MEA* acquired a new function in regulating growth during seed development. A tight regulation of the *MEA* expression level around fertilization seems to be crucial for normal development[26]. As a result, the level of overlapping *MEA* and *SWN* activity may have required adjustment after the duplication event. A pre-existing imprinting machinery may have been recruited to adjust *MEA* expression levels, leading to the recently evolved regulation of the *MEA* locus by genomic imprinting.

METHODS SUMMARY

Plant material. The *mea-1* and *swn-3* mutants of *A. thaliana* used for single- and double-mutant analyses have been previously described[4,6]. The *swn-4* mutant is the SIGNAL insertion line SALK_109121[27]. *A. lyrata* and *A. halleri* accessions were provided by T. Mitchell-Olds, M. Clauss and R. Oyama.

Expression analyses. *In situ* hybridization on embryo sac and early seed tissues was performed as previously described[5]. Riboprobes designed from the most divergent regions of the *MEA*, *SWN* and *CLF* gene sequences were used. Expression analysis of the *A. lyrata* paternal *MEA* allele in the developing seed was conducted using F_1 seed from crosses between *A. thaliana* and *A. lyrata* (where *A. lyrata* was used as a pollen parent), and quantitative polymerase chain reaction with reverse transcription (qRT–PCR) probes specific for the *MEA* transcript from each species[26].

Sequence analyses. DNA sequencing of *MEA* and *SWN* genes from *A. lyrata* and *A. halleri* was performed by high-fidelity PCR amplifications of overlapping fragments from genomic DNA and sequencing of five independent clones per PCR-amplified fragment. DNA sequence data from *A. thaliana* accessions was based on direct sequencing of high-fidelity PCR-amplified fragments.

Phylogenetic and molecular evolution analyses. Phylogenetic analysis was based on all *CLF*-, *SWN*- and *MEA*-like sequences obtained from BLAST searches of GenBank. Protein sequences were aligned with the CLUSTALW program and a phylogenetic tree constructed with the protml program of the MOLPHY package[28]. The analysis of ω-ratios was conducted with the codeml program of the PAML package[29]. Molecular population genetic analysis was conducted using the DnaSP program[30] based on sequence data from the accessions listed in Supplementary Tables 7 and 8.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 March; accepted 5 June 2007.

1. Haig, D. & Westoby, M. Parent-specific gene expression and the triploid endosperm. *Am. Nat.* **134,** 147–155 (1989).
2. Smith, F. M., Garfield, A. S. & Ward, A. Regulation of growth and metabolism by imprinted genes. *Cytogenet. Genome Res.* **113,** 279–291 (2006).
3. McVean, G. T. & Hurst, L. D. Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. *Proc. R. Soc. Lond. B* **264,** 739–746 (1997).
4. Grossniklaus, U., Vielle-Calzada, J. P., Hoeppner, M. A. & Gagliano, W. B. Maternal control of embryogenesis by *MEDEA*, a *Polycomb* group gene in *Arabidopsis*. *Science* **280,** 446–450 (1998).
5. Vielle-Calzada, J. P. *et al.* Maintenance of genomic imprinting at the *Arabidopsis medea* locus requires zygotic DDM1 activity. *Genes Dev.* **13,** 2971–2982 (1999).
6. Chanvivattana, Y. *et al.* Interaction of *Polycomb*-group proteins controlling flowering in *Arabidopsis*. *Development* **131,** 5263–5276 (2004).
7. Walter, J. & Paulsen, M. The potential role of gene duplications in the evolution of imprinting mechanisms. *Hum. Mol. Genet.* **12** (review issue 2), R215–R220 (2003).
8. Kiyosue, T. *et al.* Control of fertilization-independent endosperm development by the *MEDEA Polycomb* gene in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **96,** 4186–4191 (1999).
9. Goodrich, J. *et al.* A *Polycomb*-group gene regulates homeotic gene expression in *Arabidopsis*. *Nature* **386,** 44–51 (1997).
10. Baumbusch, L. O. *et al.* The *Arabidopsis thaliana* genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes. *Nucleic Acids Res.* **29,** 4319–4333 (2001).
11. Blanc, G., Hokamp, K. & Wolfe, K. H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13,** 137–144 (2003).
12. Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. & Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **99,** 13627–13632 (2002).
13. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
14. Springer, N. M. *et al.* Sequence relationships, conserved domains, and expression patterns for maize homologs of the *Polycomb* group genes *E(z)*, *esc*, and *E(Pc)*. *Plant Physiol.* **128,** 1332–1345 (2002).
15. Springer, N. M. *et al.* Comparative analysis of SET domain proteins in maize and *Arabidopsis* reveals multiple duplications preceding the divergence of monocots and dicots. *Plant Physiol.* **132,** 907–925 (2003).
16. Mayama, T., Ohtsubo, E. & Tsuchimoto, S. Isolation and expression analysis of petunia *CURLY LEAF*-like genes. *Plant Cell Physiol.* **44,** 811–819 (2003).

17. Thakur, J. K. *et al.* A *Polycomb* group gene of rice (*Oryza sativa* L. subspecies indica), *OsiEZ1*, codes for a nuclear-localized protein expressed preferentially in young seedlings and during reproductive development. *Gene* **314,** 1–13 (2003).
18. Ohad, N. *et al.* A mutation that allows endosperm development without fertilization. *Proc. Natl Acad. Sci. USA* **93,** 5319–5324 (1996).
19. Chaudhury, A. M. *et al.* Fertilization-independent seed development in *Arabidopsis thaliana. Proc. Natl Acad. Sci. USA* **94,** 4223–4228 (1997).
20. Guitton, A. E. *et al.* Identification of new members of *FERTILISATION INDEPENDENT SEED Polycomb* group pathway involved in the control of seed development in *Arabidopsis thaliana. Development* **131,** 2971–2981 (2004).
21. Köhler, C. *et al. Arabidopsis* MSI1 is a component of the MEA/FIE *Polycomb* group complex and required for seed development. *EMBO J.* **22,** 4804–4814 (2003).
22. Wang, D., Tyson, M. D., Jackson, S. S. & Yadegari, R. Partially redundant functions of two SET-domain *Polycomb*-group proteins in controlling initiation of seed development in *Arabidopsis. Proc. Natl Acad. Sci. USA* **103,** 13244–13249 (2006).
23. Yang, Z. & Bielawski, J. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).
24. Bielawski, J. P. & Yang, Z. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.* **59,** 121–132 (2004).
25. Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weishaar, B. & Mitchell-Olds, T. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169,** 1601–1615 (2005).
26. Baroux, C., Gagliardini, V., Page, D. R. & Grossniklaus, U. Dynamic regulatory interactions of *Polycomb* group genes: *MEDEA* autoregulation is required for imprinted gene expression in *Arabidopsis. Genes Dev.* **20,** 1081–1086 (2006).
27. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana. Science* **301,** 653–657 (2003).
28. Adachi, J. H. M. MOLPHY Version 2.3: Programs for molecular phylogenetics based on Maximum Likelihood. *Comp. Sci. Monogr.* **28,** 1–150 (1996).
29. Yang, Z. & Nielsen, R. Synonymous and non-synonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46,** 409–418 (1998).
30. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. DNAsp, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19,** 2496–2497 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Gheyselinck and P. Kopf for the technical support; C. O'Mahony for assistance with artwork and figures; M. O' Connell for comments on the manuscript; and T. Mitchell-Olds, M. Clauss, R. Oyama, J. Goodrich and NASC for seeds. This work was supported by the University of Zürich, a UNESCO fellowship (to J.-M.E.-R.), the EU Network of Excellence 'EPIGENOME', and grants of the Swiss National Science Foundation (to U.G.), the Deutsche Forschungsgemeinschaft and the Max Planck Society (to K.J.S.), and the Science Foundation Ireland (to C.S. and K.H.W.).

Author Information Sequences generated in this study are available from GenBank (accession numbers DQ975464–DQ975465). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to U.G. (grossnik@botinst.uzh.ch) or C.S. (c.spillane@ucc.ie).

METHODS

Germplasm, DNA sequencing, mutants and crosses. The *mea-1* mutant (L*er*-0) used in this study contains a *Ds* insertion in the SET domain of the *MEA* gene (At1g02580) and has been previously described[4]. Heterozygous or homozygous lines of *mea-1* were identified by PCR genotyping using the primer combinations Ds5-1/AS13 (*mea-1* present) and S20/AS13 (*MEA* present) as previously described. The *swn/eza1* allele (*swn-3*) used for the single- and double-mutant analysis of the post-fertilization seed abortion phenotype was a mutant line (SALK_050195 in Col-0 background) from the SIGNAL collection[27] that contains an insertion in exon 15 within the SET domain of the *EZA1*/*SWN* gene[6]. The *swn-4* allele used corresponds to the SIGNAL insertion line SALK_109121 where the T-DNA insertion is located in exon 8 of the *SWN* gene. The *clf-2* allele was obtained from J. Goodrich. To construct double mutants of *mea-1* and *swn-3*, pollen from a *mea-1* homozygous line was used to pollinate *swn-3* heterozygous plants and the F_1 seed progeny selected on MS medium containing kanamycin. F_1 progeny, which were double heterozygotes (*mea-1/MEA; swn-3/SWN*), were identified by PCR genotyping and selfed to generate F_2 progeny segregants. PCR genotyping was used to identify genotypes among the F_2 segregants. The *A. lyrata* (Bish Bash) accession used for crosses and sequencing of the *MEA* and *SWN* orthologues was provided by T. Mitchell-Olds. The *A. lyrata* and *A. halleri* accessions for molecular population genetic studies were provided by M. Clauss and R. Oyama.

***In situ* hybridization.** *In situ* hybridization was performed as described[5,31,32] with modifications. Mature flowers and siliques of *A. thaliana* plants were fixed in 4% paraformaldehyde and embedded in Paraplast Plus (Sigma). Sections of 10-μm thickness were cut with a Leica microtome (Leica RM 2145) mounted on ProbeOnPlus slides (Fischer Biotech). Sections were digested with proteinase K for 30 min at 37 °C, treated with acetic acid anhydride, dried in ethanol, then hybridized with 11-digoxigenin-UTP (DIG)-labelled probes overnight at 55 °C. After washing with 0.2× SSC at 55 °C, the slides were processed for revealing the DIG antigen. This involved blocking with DIG-blocking reagent and BSA, followed by incubation with an ant-DIG antibody conjugated to alkaline phosphatase (Roche Diagnostics), washing with blocking reagent, then colour revealed by incubation in NBT and X-phosphate for periods of 16 to 18 h. Reactions were stopped with TE buffer (10 mM, pH 8.0), then mounted in TE/glycerol (1 : 4 v/v) before viewing. The riboprobes used for *in situ* hydridization were synthesized from RT–PCR products using primers designed on sequence data available. The genes were *CLF* (At2g23380), *SWN* (At4g02020) and *MEA* (At1g02580). These probes were designed to cover the coding region of the genes analysed and to avoid cross hybridization. For synthesis of sense and anti-sense DIG-labelled probes, 350-bp fragments have been cloned in pBluescript SK- vector for each of the genes analysed. For hybridization probe design, the most divergent regions of *MEA*, *SWN* and *CLF* were identified by ClustalX alignment of the coding sequence of each gene. The divergent regions chosen for *in situ* probe construction were: *SWN* (355–606 bp downstream of start codon in messenger RNA); *MEA* (608–955 bp downstream of start codon in mRNA); and *CLF* (33–364 bp downstream of start codon in mRNA). The primer pairs used for amplification of these regions and cloning into the expression vector (pBluescript SK-) for riboprobe synthesis were: *SWN* (*SWN* exon 4F, 5′-GCAGAAATTTGAGGCTAATAG-3′ and *SWN* exon 4R, 5′-CCAGGTAGTGTATGGCGG-3′); *MEA* (*MEA in situ* F, 5′-CGGTTGGGCAGGACTATGG-3′ and *MEA in situ* R, 5′-CTTCTGTCACACTCCTCACC- 3′); *CLF* (*CLF in situ* F, 5′-CACCAGATCGGAGCCACC-3′ and *CLF in situ* R, 5′-GACAGGGACACTAGATCC-3′). Reverse transcription was performed using AMV reverse transcriptase (Clontech) and total RNA (1 μg) extracted from *A. thaliana* siliques using Triazol (GIBCO-BRL).

Expression analysis of *A. lyrata* paternal *MEA* allele in developing seeds. Crosses were conducted between *A. thaliana* and *A. lyrata* as a pollen parent. RNA was extracted at specific time points before or after pollination using trizol, and the accumulation of *MEA* transcripts was measured using quantitative real-time RT–PCR as previously described[26]. Quantitative analyses of transcript levels were carried out using Taqman real-time PCR assays (Applied Biosystem). Three PCR replicates were performed for each cDNA sample, and the specificity and amount of the unique amplification product were determined according to the manufacturer's instructions (Applied Biosystems). To distinguish between maternal (*A. thaliana*) and paternal (*A. lyrata*) *MEA* transcripts, we used probes that specifically recognize the different alleles. In all experiments, transcript levels were normalized to the level of *ACTIN-11*, which is expressed in the gametophyte and zygotic products of the seed (embryo and endosperm) but not in the surrounding maternal tissues[33]. Beyond 4 days after pollination (d.a.p.), *ACTIN-11* levels decrease and cannot be used for normalization (data not shown). The primers used for the real-time assay were: (1) for detection of the *MEA* allele from *A. thaliana* (*MEA-At*), forward 5′-TCTGATGTTCATGGATGGGG-3′; reverse 5′-GGTAGGAAGAACCAATCCGATCT-3′; probe VIC 5′-TCACTCATGATGAAGCTAA-3′ MGB (ABI); (2) for detection of the *MEA* allele from *A. lyrata* (*MEA-Al*), forward 5′-ATCAAGGTTGTGTTTTTAATAAAGAGGC-3′; reverse 5′-CAGCTGGCTACTTTTGATGAAGAC-3′; probe FAM 5′-ACCTTCCAGTTGTTGAGC-3′ MGB (ABI).

DNA sequencing of *MEA* and *SWN* orthologues in *A. lyrata* and *A. halleri*. The sequences of the *MEA* and *SWN* genes from *A. lyrata* (Bish Bash) were initially determined by high fidelity PCR amplification of overlapping fragments from genomic DNA and sequencing of five independent clones (in pGEM) per PCR-amplified fragment. The primer pairs used were designed to be specific to either the *MEA* or *SWN* genes. The *A. lyrata* and *A. halleri* accessions used for sequencing of *MEA* are indicated in Supplementary Table 8. The following overlapping primer pairs were used for amplification of the *MEA* gene from *A. lyrata* and *A. halleri* accessions. *MEA* ORF: MEA-F1/MEA-R1, MEA-F2/MEA-R2, MEA-F3/MEA-R3, MEA-F4/MEA-R4, MEA-F5/MEA-R5, MEA-PF/MEA-PR, MEA-P1/MEA-P2, MEA-P3/MEA-P4. The following overlapping primer pairs were used to amplify the *SWN* ORF from *A. lyrata*: SWN-F1/SWN-R1, SWN-F2/SWN-R2, SWN-F3/SWN-R4, SWN-F4/SWN-R5. The sequences of the primers are listed in Supplementary Table S9.

Phylogenetic analysis. The phylogenetic analysis was based on all *CLF*-, *MEA*- and *SWN*-like sequences obtained from BLAST searches of GenBank. The protein and gene IDs used for the tree construction were: *A. thaliana*, MEA (NP_563658/NM_100139), SWN/EZA1 (AAL90954/AY090293), CLF (AAC23781/AC003040); *A. lyrata*, MEA (bankit835839), SWN/EZA1 (bankit842314); *Zea mays*, MEZ1 (AAM13420/AF443596), MEZ2 (AAM13421/AF443597), MEZ3 (AAM13422/AF443598); *Oryza sativa*, SET1 (AAN01115/AF407010), EZ1 (*O. s. indica*) (CAD18871/AJ421722), CLF (*O. s. japonica*) (NP_910690/NM_185801), EZ1 (*O. s. japonica*) (BAD69169/AP005813); Petunia x Hybrida, PHCLF1 (BAC84950/AB098523), PHCLF2 (BAC84951/AB098524), PHCLF3 (BAC84952/AB098525). The poplar sequences were obtained by searching the poplar genome database at the JGI (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html) (protein IDs: EZA1-like1, 731686; EZA1-like2, 348349; CLF-like1, 719252; CLF-like2, 694432). Protein sequences were aligned with the CLUSTALW program and a phylogenetic tree was constructed with the protml program of the MOLPHY package, which implements a maximum likelihood method[28]. The 'quick add OTUs search' strategy was used in with the JTT substitution matrix and the six best trees were retained. Subsequently, each tree was swapped and re-optimized (local rearrangement search), and branch lengths and local bootstrap probabilities (LBP) were estimated during this last search, leading to the tree with the highest likelihood shown in Fig. 1c.

Evolutionary analysis and tests of natural selection. The ratio $\omega = d_N/d_S$—with d_N as the number of non-synonymous substitutions per non-synonymous site and d_S as the number of synonymous substitutions per synonymous site—was used to test whether protein-coding sequences evolve under positive darwinian selection[34]. The analyses were carried out with the codeml program of the PAML package[29], which uses maximum likelihood estimation of parameters. To test for positive or purifying selection, ω-ratios for different site classes of the coding sequence were estimated. The likelihood of this estimate was then compared with the likelihood of other models with different numbers of parameters. A likelihood ratio test was applied by calculating the test statistic as the twofold difference of the two likelihoods ($2\delta = 2(l_1 - l_2)$); the critical values were looked up in a χ^2 table with the degrees of freedom calculated as the difference of the parameters that were estimated by each model. Two types of models were analysed. Branch models allow different ω-ratios in different branches of the phylogeny and can be used to address whether selection pressures on proteins are variable among species or among paralogues of a gene family. We used branch models with one (Model M0), two, three, four or n (the total number of branches; free-ratio model) ω-ratios[35]. The second type of models analysed were branch-site models[34]. These models allow one to test whether positive selection occurred in a subset of codons in a particular branch of the phylogeny (the 'foreground branch') by assuming two types of codons with $0 < \omega < 1$ and $\omega = 1$ in the entire tree and an additional class of codons with $\omega > 1$ in the foreground branch. After estimating ω-ratios, a bayesian empirical Bayes algorithm was applied to identify amino acid residues with a high posterior probability of $\omega > 1$. Among available branch-site models, we used model A of ref. 35 to test whether functional diversification of newly duplicated genes was driven by positive selection.

Sequence analysis of divergent accessions of *A. thaliana*, *A. lyrata* and *A. halleri*. Sequences of *SWN* and *MEA* genes were amplified from 21 divergent accessions of a worldwide collection of *A. thaliana* (Supplementary Table 7) using overlapping PCR primers (Supplementary Table 9). PCR products were directly sequenced on an ABI 3730xl automated sequencer using dye terminator chemistry. Sequence data were assembled and aligned with an automated sequence analysis pipeline as described[36]. The *MEA* gene was also obtained from

nine individuals originating from geographically distant populations of *A. lyrata* and from ten individuals from the close relative *A. halleri* (Supplementary Table 8). High fidelity PCR amplifications were performed using the Phusion High-Fidelity PCR Kit (FINNZYMES) and the overlapping PCR primers described for the *A. lyrata* sequencing. A first set of PCR amplifications was directly sequenced, and a second set of independent PCR amplifications was cloned before sequencing. Briefly, following an A-tailing step, the PCR products were cloned into the pGEM-T easy vector following the manufacturer's instructions (Promega). For each PCR fragment, two independent clones were sequenced in both directions (Macrogen Inc.). All sequences generated for the *MEA* gene were analysed using the DNASTAR software package (DNASTAR). All polymorphisms were inspected visually. Molecular population genetic analysis was carried out with the DnaSP program[30]. Nucleotide diversity was calculated as the average pairwise nucleotide diversity, π_{tot}, and haplotype diversity as H_d[37]. Several tests of neutral evolution using the polymorphism data where applied. Tajima's D[38] tests whether there is an excess of low- or high-frequency polymorphisms and was calculated with silent (synonymous coding and non-coding) polymorphisms; the H statistic of ref. 39 analyses the frequency spectrum of derived polymorphisms and was also calculated with silent polymorphisms; the McDonald–Kreitman test[40] compares the ratio of non-synonymous to synonymous polymorphisms and fixed differences; the Hudson–Kreitman–Aguade test was used to test whether the polymorphism to divergence ratio between the two genes was significantly different from each other, which is expected if selection acts on one gene but not on the other[41].

31. Coen, E. S. *et al. floricaula*: a homeotic gene required for flower development in *Antirrhinum majus*. *Cell* **63,** 1311–1322 (1990).
32. Jackson, D., Culianez-Macia, F., Prescott, A. G., Roberts, K. & Martin, C. Expression patterns of *myb* genes from *Antirrhinum* flowers. *Plant Cell* **3,** 115–125 (1991).
33. Huang, S., An, Y. Q., McDowell, J. M., McKinney, E. C. & Meagher, R. B. The *Arabidopsis ACT11* actin gene is strongly expressed in tissues of the emerging inflorescence, pollen, and developing ovules. *Plant Mol. Biol.* **33,** 125–139 (1997).
34. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17,** 32–43 (2000).
35. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19,** 908–917 (2002).
36. Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B. & Mitchell-Olds, T. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169,** 1601–1615 (2005).
37. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
38. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
39. Fay, J. C. & Wu, C.-I. Hitchhiking under positive Darwinian selection. *Genetics* **155,** 1405–1413 (2000).
40. McDonald, J. & Kreitman, M. Adaptive evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
41. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).

LETTERS

Variants conferring risk of atrial fibrillation on chromosome 4q25

Daniel F. Gudbjartsson[1], David O. Arnar[2], Anna Helgadottir[1], Solveig Gretarsdottir[1], Hilma Holm[2], Asgeir Sigurdsson[1], Adalbjorg Jonasdottir[1], Adam Baker[1], Gudmar Thorleifsson[1], Kristleifur Kristjansson[1], Arnar Palsson[1], Thorarinn Blondal[1], Patrick Sulem[1], Valgerdur M. Backman[1], Gudmundur A. Hardarson[1], Ebba Palsdottir[1], Agnar Helgason[1], Runa Sigurjonsdottir[2], Jon T. Sverrisson[3], Konstantinos Kostulas[4], Maggie C. Y. Ng[5], Larry Baum[5], Wing Yee So[5], Ka Sing Wong[5], Juliana C. N. Chan[5], Karen L. Furie[6], Steven M. Greenberg[6], Michelle Sale[6], Peter Kelly[6], Calum A. MacRae[7], Eric E. Smith[6], Jonathan Rosand[6], Jan Hillert[4], Ronald C. W. Ma[5], Patrick T. Ellinor[7], Gudmundur Thorgeirsson[2], Jeffrey R. Gulcher[1], Augustine Kong[1], Unnur Thorsteinsdottir[1] & Kari Stefansson[1]

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia in humans and is characterized by chaotic electrical activity of the atria[1]. It affects one in ten individuals over the age of 80 years, causes significant morbidity and is an independent predictor of mortality[2]. Recent studies have provided evidence of a genetic contribution to AF[3–5]. Mutations in potassium-channel genes have been associated with familial AF[6–10] but account for only a small fraction of all cases of AF[11,12]. We have performed a genome-wide association scan, followed by replication studies in three populations of European descent and a Chinese population from Hong Kong and find a strong association between two sequence variants on chromosome 4q25 and AF. Here we show that about 35% of individuals of European descent have at least one of the variants and that the risk of AF increases by 1.72 and 1.39 per copy. The association with the stronger variant is replicated in the Chinese population, where it is carried by 75% of individuals and the risk of AF is increased by 1.42 per copy. A stronger association was observed in individuals with typical atrial flutter. Both variants are adjacent to *PITX2*, which is known to have a critical function in left–right asymmetry of the heart[13–15].

We conducted a genome-wide association study with the use of the Illumina Hap300 BeadChip on an Icelandic population with AF and/or atrial flutter (AFl); 316,515 single-nucleotide polymorphisms (SNPs) satisfying our quality criteria (Supplementary Information) were tested individually for association with AF or AFl in a sample of 550 patients and 4,476 controls from Iceland. Three strongly correlated SNPs, all located within a single linkage disequilibrium (LD) block on chromosome 4q25, were the only SNPs found to be significant on a genome-wide basis after the 316,515 SNPs tested had been accounted for ($P < 0.05/316{,}515 = 1.58 \times 10^{-7}$): rs2200733 (odds ratio (OR) = 1.75; $P = 1.6 \times 10^{-10}$), rs2220427 (OR = 1.75; $P = 1.9 \times 10^{-10}$) and rs2634073 (OR = 1.60; $P = 2.1 \times 10^{-9}$). These results and all other results based on the Icelandic population were adjusted for the relatedness of individuals. The two most significant SNPs, rs2200733 and rs2220427, are perfect proxies for one another in the CEPH CEU HapMap[16] data set and are close to being perfect proxies for one another in the Icelandic data set ($D' = 1$, $r^2 = 0.999$); therefore only rs2200733 will be referred to in the following discussion. The correlation of rs2634073 with rs2200733 is weaker in the Icelandic data set ($D' = 0.95$, $r^2 = 0.605$). On further study of the Illumina Hap300 SNPs in the vicinity of the first three SNPs and conditioning on the association with rs2200733, an association with a new SNP, rs10033464, was identified (OR = 1.42; $P = 0.0024$). After the association with rs2200733 and rs10033464 had been accounted for, the association with rs2634073 was no longer significant ($P = 0.30$). Henceforth, all association results for rs2200733 T and rs10033464 T, including those presented in Table 1, are based on a comparison with the wild-type haplotype, which carries neither of the two at-risk alleles, rather than on a comparison with the major alleles of each SNP separately. Specifically, ORs for rs2200733 T and rs10033464 T are each computed conditionally (see Methods Summary) and could be interpreted as the estimated relative risk of each variant compared with the wild type. The at-risk alleles T of rs2200733 and T of rs10033464 have estimated population allelic frequencies of 12.05% and 8.53% in Iceland, respectively, and are never observed together on the same chromosome, in either the Icelandic data set or the CEU HapMap data set. A third SNP, rs13143308, which has a minor allele that corresponds completely to chromosomes carrying either the T allele of rs2200733 or the T allele of rs10033464, was identified through the CEU HapMap data set. Figure 1 demonstrates the haplotype structure over the key SNPs of the associated region. Sets of SNPs that are perfect proxies of each of these three key SNPs in the CEU HapMap samples are provided in Supplementary Table 1, and relative locations are shown in Fig. 2. We emphasize that the SNPs named should be considered representatives of the haplotypes defined by the SNPs to which they are equivalent and are chosen primarily for the sake of convenience.

A microsatellite marker, D4S406, located in the same LD block as the two SNPs was identified. In Iceland, three of the four shortest alleles of D4S406 (−8, −4 and −2) combine to form a near-perfect surrogate for the T allele of rs2200733 ($D' = 0.995$, $r^2 = 0.98$), and the two shortest remaining alleles (−6 and 0) form a good surrogate for the T allele of rs10033464 ($D' = 0.98$, $r^2 = 0.75$; Supplementary Table 2). None of the remaining (longer) alleles of D4S406 are associated with AF/AFl after the effect of the short alleles had been accounted for. For replication of the original observation in

[1]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland. [2]Division of Cardiology, Department of Medicine, Landspitali University Hospital, 101 Reykjavik, Iceland. [3]Department of Medicine, Akureyri Regional Hospital, 600 Akureyri, Iceland. [4]Department of Neurology, Karolinska Institutet at Karolinska University Hospital, Huddinge S-141 86, Sweden. [5]Department of Medicine and Therapeutics, Prince of Wales Hospital, Chinese University of Hong Kong, Shatin, Hong Kong. [6]Department of Neurology, [7]Cardiology Division and Cardiovascular Research Center, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA.

Table 1 | Analysis of the association of rs2200733 and rs10033464 on chromosome 4q25 with AF/AFl

Sample (cases/controls)	rs2200733 T* Frequency§	OR (95% CI)	P	rs10033464 T*† Frequency§	OR (95% CI)	P	Comparison P‡	Joint PAR
Iceland‖								
Discovery (550/4,476)	0.191 (0.114)	1.84 (1.54–2.21)	2.0×10^{-11}	0.110 (0.080)	1.42 (1.13–1.77)	0.0024	0.041	0.216
Replication (2,251/13,238)	0.166 (0.108)	1.64 (1.49–1.81)	2.7×10^{-23}	0.108 (0.080)	1.40 (1.24–1.58)	8.2×10^{-8}	0.028	0.176
Combined (2,801/17,714)	0.171 (0.110)	1.68 (1.53–1.83)	1.9×10^{-30}	0.108 (0.080)	1.40 (1.25–1.55)	9.4×10^{-9}	0.0025	0.180
Other European ancestry								
Sweden (143/738)	0.179 (0.098)	2.01 (1.38–2.93)	0.00027	0.172 (0.111)	1.65 (1.14–2.41)	0.0087	0.41	0.272
United States (636/804)	0.229 (0.139)	1.84 (1.51–2.23)	9.8×10^{-10}	0.105 (0.083)	1.30 (1.00–1.69)	0.052	0.026	0.232
Combined¶	– (–)	1.88 (1.58–2.23)	1.2×10^{-12}	– (–)	1.41 (1.13–1.75)	0.0019	0.027	0.237
All European ancestry								
Combined¶	– (–)	1.72 (1.59–1.86)	3.3×10^{-41}	– (–)	1.39 (1.26–1.53)	6.9×10^{-11}	0.00019	0.206
Hong Kong								
Hong Kong (333/2,836)	0.605 (0.528)	1.42 (1.16–1.73)	0.00064	0.190 (0.218)	1.08 (0.84–1.39)	0.55	0.0099	0.346

Each row contains the results from a joint analysis of two variants, rs2200733 T and rs10033464 T†. The numbers of cases and controls are shown for each case-control study and for each variant the allelic frequencies of the variant in cases and controls, the OR with a 95% CI and two-sided *P* values are shown. In addition a *P* value for comparing the effect of the two variants and their joint PAR is reported. For example, the first row indicates that, for the initial Icelandic discovery samples, rs2200733 T has an estimated OR of 1.84 (95% CI 1.54–2.21, $P = 2.0 \times 10^{-11}$) versus the wild type (rs2200733 C, rs13143308 G, rs10033464 G haplotype), and rs2200733 T has an estimated OR of 1.42 (95% CI 1.13–1.77, $P = 0.0024$) versus the wild type.

* Results of comparing rs2200733 T and rs10033464 T with the wild-type rs2200733 C, rs13143308 G, rs10033464 G haplotype.

† In the Swedish and US samples rs10033464 T was tagged by the rs2200733 C, rs13143308 T haplotype.

‡ *P* value for comparing the ORs of rs2200733 T and rs10033464 T.

§ Shown as cases (controls).

‖ The association analysis was adjusted for the relatedness of some of the individuals.

¶ For the combined study populations of European decent, the PAR was calculated by using the average, unweighted control frequency of the populations, whereas the OR and *P* value were estimated by using the Mantel–Haenszel model.

Iceland the D4S406 genotypes were used to provide information when SNP genotypes were not available.

In an attempt to replicate our original discovery we analysed an additional Icelandic sample consisting of 2,251 AF/AFl patients and 13,238 controls (Table 1). The association of both SNPs with AF/AFl was replicated in this sample (OR = 1.64, $P = 2.7 \times 10^{-23}$ for rs2200733; OR = 1.40, $P = 8.2 \times 10^{-8}$ for rs10033464) and both achieve genome-wide significance in the combined Icelandic samples (OR = 1.68, $P = 1.9 \times 10^{-30}$ for rs2200733; OR = 1.38, $P = 9.4 \times 10^{-9}$ for rs10033464). We also typed all the 18 Hap300 Illumina SNPs in the region around our signal in 404 of the additional AF cases and 2,036 of the additional controls. None of these SNPs remained significant after the association with rs2200733 and rs10033464 had been accounted for (Supplementary Table 3).

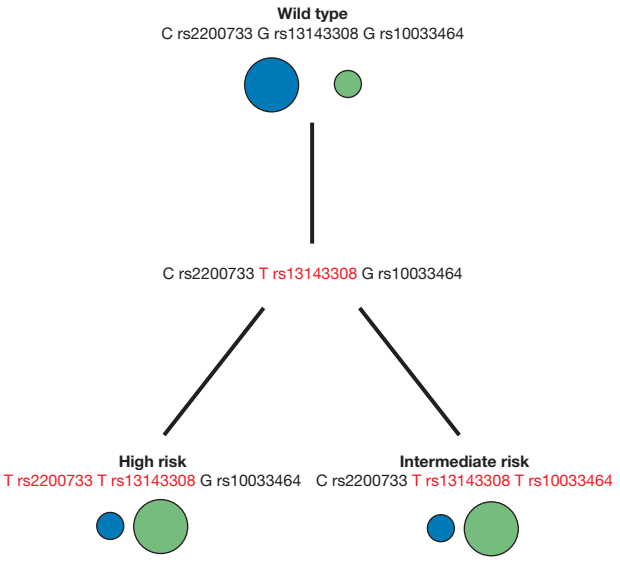


Figure 1 | Diagram of the haplotype structure at the associated region. Each edge in the graph corresponds to one mutation. The areas of the blue circles are proportional to the haplotype frequencies of the haplotypes in Iceland, and the areas of the green circles are proportional to the haplotype frequencies in Hong Kong. Note that the intermediary haplotype, shown in the middle of the graph, no longer exists with certainty in either of the two populations (its estimated frequency is less than 0.2% which is indistinguishable from genotyping errors).

In further attempts to replicate our results, we tested these variants for an association with AF in two populations of European ancestry, one from Sweden, consisting of 143 cases and 738 controls, and the other from the United States, consisting of 636 cases and 804 controls (Table 1). The association with rs2200733 was strongly replicated in both populations (OR = 2.01, $P = 0.00027$ in Sweden; OR = 1.84, $P = 9.8 \times 10^{-10}$ in the United States). The association with rs10033464 is weaker but was nonetheless replicated in the Swedish population (OR = 1.65, $P = 0.0087$) and was nearly significant in the US population (OR = 1.30, $P = 0.052$). When combined with the Icelandic samples, the association with rs2200733 was unequivocal (OR = 1.72, $P = 3.3 \times 10^{-41}$), and the significance of rs10033464 was well beyond the threshold of genome-wide significance (OR = 1.39, $P = 6.9 \times 10^{-11}$). If the multiplicative model is assumed, the population attributable risk (PAR) of the two variants combined is about 20% in populations of European ancestry.

Finally, we attempted to replicate these signals in a Han Chinese population from Hong Kong consisting of 333 cases and 2,836 controls. The association with rs2200733 T was significantly replicated (OR = 1.42, $P = 0.00064$), but the association with rs10033464 T was not significant, although the direction of association was consistent with that in the European samples (OR = 1.08, $P = 0.55$; Table 1). The T allele of rs2200733 is much more frequent in the Chinese population (the allelic frequency in controls is 0.528) than in those of European descent (allelic frequency in controls 0.098–0.139; Fig. 1), which is reflected in a greater joint PAR of about 35%, even though the estimated risk is less. The LD block containing the two variants is more fragmented in the Chinese CHB and Japanese JPT HapMap samples than in the CEU HapMap samples (Fig. 2). We therefore analysed several markers in the Hong Kong population that were in perfect LD with rs2200733 in the CEU samples but in imperfect LD in the CHB and JPT samples (Supplementary Table 4). These markers had a weaker apparent association with AF than with rs2200733, suggesting that the functional variants driving the association is located in the roughly 20-kilobase (kb) region around the original rs2200733 variant and defined by the SNPs that remain equivalent to rs2200733 in the CHB and JPT samples (red in Fig. 2).

For the initial Icelandic discovery samples, rs2200733 had a significantly higher OR than rs10033464 ($P = 0.041$). This held true in the replication samples, and overall there is a significant difference in the risks associated with the two variants ($P = 0.00019$ in the combined European samples and $P = 0.0099$ in Hong Kong). When genotype-specific ORs were studied, some deviation from the multiplicative model is detectable in the combined data set ($P = 0.018$ for

European samples; see Supplementary Table 5). Estimated risks of heterozygous carriers relative to non-carriers were similar, but homozygous carriers of rs2200733 T and rs10033464 T have estimated risks that were, respectively, higher and lower than predicted by a multiplicative model. A similar trend was seen in the Hong Kong samples, although the sample size was too small to have power to detect such deviations with significance. In the combined populations of European descent, the observed OR for individuals homozygous for rs2200733 T was 3.64 in comparison with individuals homozygous for the wild-type haplotype, and 1.77 for the Chinese population, showing that these variants are important components in any predictive modelling of AF.

The age at diagnosis of AF/AFl for the Icelandic samples is correlated with the two SNPs (diagnosis occurs 2.28 years earlier per T allele of rs2200733 and 1.10 years earlier per T allele of rs10033464; joint $P = 1.29 \times 10^{-6}$). The effect of the age at diagnosis was also evaluated by measuring the strength of association while stratifying by age at diagnosis. The association of the two variants is strongest in those diagnosed at a younger age, although the risk remains significant even in those diagnosed after reaching 80 years of age (Table 2). Information on age at diagnosis of AF was not available for the Swedish samples. The US samples consisted of two main groups: younger patients with either lone AF or AF and hypertension, and older AF cases who were mostly patients with haemorrhagic and ischaemic stroke. In both populations there is a clear trend towards a stronger association in younger AF cases than in older cases. Our analysis of the data did not suggest any differential association by sex (Table 2).

AF1 often accompanies AF, but it can occur in isolation[17]; we observed a strong association between the variants and the small subset ($N = 116$) of the AF1 Icelandic patients (OR = 2.60, 95% confidence interval (CI) = 1.83–3.68, $P = 7.5 \times 10^{-8}$ for rs2200733; OR = 1.94, 95% CI = 1.26–3.00, $P = 0.0028$ for rs10033464). Indeed, for rs2200733, the OR for these definite AFl cases is significantly higher than that for the cases with an AF phenotype ($P = 0.0026$), and is close to significantly higher for rs10033464 ($P = 0.084$). Our results suggest that while these traits share genetic risk factors, AFl is less influenced by phenocopies than AF.

Neither variant showed an association with obesity, hypertension or myocardial infarction in the Icelandic samples, all known risk factors for AF (observed OR < 1.1 in all instances; Supplementary Table 6). Although these negative results do not exclude the possibility that the new variants associate with these phenotypes, they do suggest, along with the high risk in US lone AF and earlier age at onset in carriers, that the new variants are not affecting the risk of AF through these known risk factors.

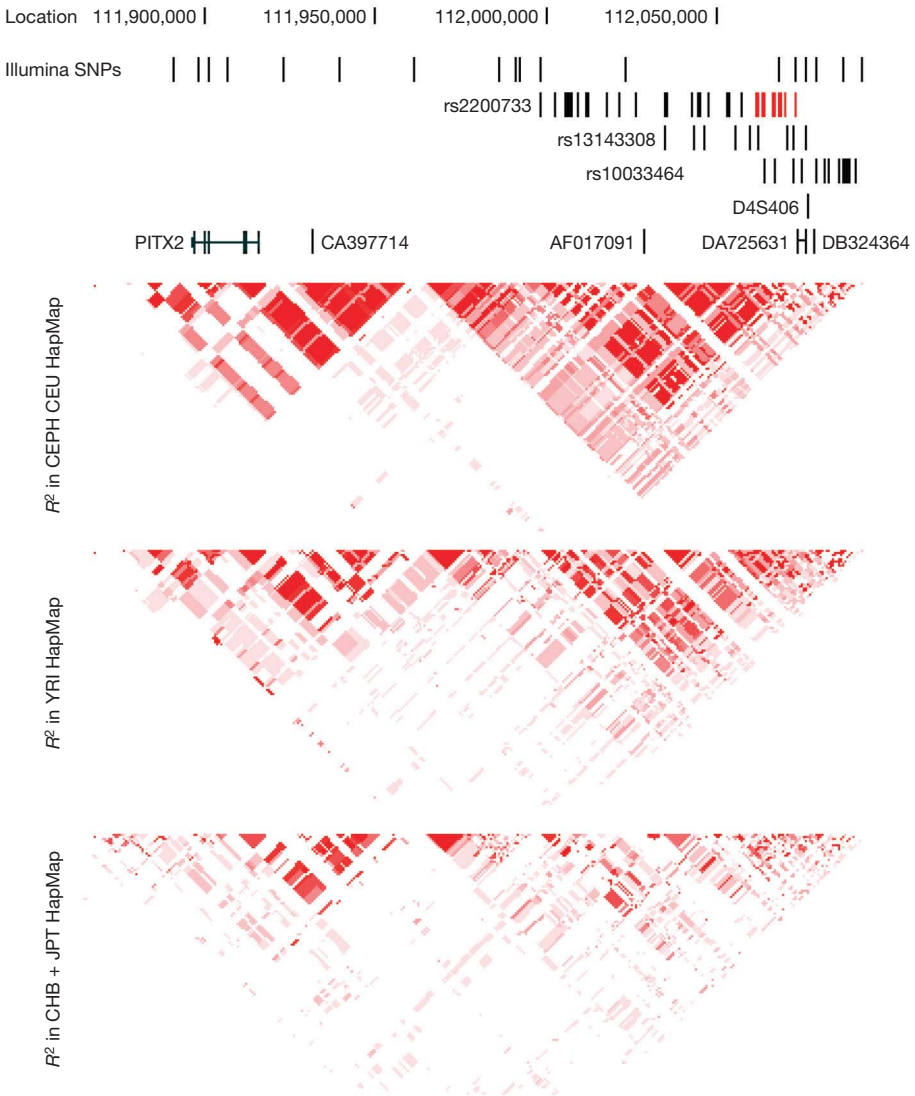


Figure 2 | Overview of a 200-kb genomic neighbourhood of rs2200733 and rs10033464. Predicted ESTs, the locations of the three main classes of equivalent SNPs in the CEU HapMap samples and an overview of the LD structure of the region in the CEU, CHB and JPT HapMap samples are shown. The SNPs that remain equivalent to rs2200733 in the CHB and JPT HapMap samples are shown in red.

Table 2 | Association by age at diagnosis in Iceland and by AF sub-phenotype in the United States

Sample (cases/controls)	Male (%)	Age (yr)	OR (95% CI)		P	Sex P
			rs2200733*	rs10033464*†		
Iceland‡						
Diagnosis at age ≤60 yr (510/17,714)	77.8	50.7 ± 8.4	2.12 (1.77–2.54)	1.69 (1.34–2.12)	6.3×10^{-18}	0.82
Diagnosis at age 60–70 yr (654/17,714)	66.2	65.6 ± 2.9	1.88 (1.60–2.21)	1.44 (1.18–1.77)	6.7×10^{-15}	0.58
Diagnosis at age 70–80 yr (958/17,714)	58.9	75.0 ± 2.8	1.60 (1.39–1.84)	1.23 (1.03–1.47)	7.5×10^{-11}	0.96
Diagnosis at age >80 yr (679/17,714)	47.4	85.6 ± 4.2	1.20 (1.01–1.43)	1.31 (1.08–1.60)	0.0044	0.36
United States						
Lone AF (251/804)	81.7	46.1 ± 11.5	2.32 (1.80–2.99)	1.68 (1.19–2.37)	1.2×10^{-10}	0.46
AF + hypertension (67/804)	74.6	54.5 ± 10.2	2.23 (1.43–3.48)	1.66 (0.90–3.04)	0.0010	0.54
Other AF (318/804)	52.8	75.2 ± 11.3	1.44 (1.12–1.84)	0.97 (0.69–1.37)	0.015	0.85

Each row contains the results from a joint analysis of two variants, rs2200733 T and rs10033464 T*. The numbers of cases and controls, the percentage of male cases and the age (mean ± s.d.) for cases are shown for each case-control study. The OR, with a 95% CI, and *P* values are shown for each variant. In addition a joint *P* value is shown for the combined effect of the two variants, as is a joint *P* value for testing whether there is a difference in the allelic frequency of the variants between the sexes within each subgroup of patients.
* Results of comparing rs2200733 T and rs10033464 T with the wild-type rs2200733 C, rs13143308 G, rs10033464 G haplotype.
† In the US samples, rs10033464 T was tagged by the rs2200733 C, rs13143308 T haplotype.
‡ The association analysis was adjusted for the relatedness of some of the individuals.

There is no known gene present in the LD block containing rs2200733 and rs10033464 (Fig. 2). The LD block contains one spliced expressed sequence tag (EST) (DA725631) and two single-exon ESTs (DB324364 and AF017091). Reverse-transcriptase-mediated polymerase chain reaction of complementary DNA libraries from various tissues did not detect the expression of these ESTs (Supplementary Information). The *PITX2* gene located in the adjacent upstream LD block is the gene closest to the risk variants. The protein encoded by this gene, the paired-like homeodomain transcription factor 2, is an interesting candidate for AF/AFl because it is known to be important in cardiac development by directing the asymmetric morphogenesis of the heart[13]. In a mouse knockout model, Pitx2 was shown to suppress a default pathway for sinoatrial node formation in the left atrium[14,15]. There is very little mRNA expression of *PITX2* in all easily accessible tissues, such as blood and adipose tissue, hampering the study of correlation between genotypes and expression levels. The next gene upstream of *PITX2* is *ENPEP*, an aminopeptidase responsible for the breakdown of angiotensin II in the vascular endothelium[18]. This gene is expressed more widely but the variants associated with AF showed no correlation with its expression in blood or adipose tissue (Supplementary Information). No other annotated genes are located within a 400-kb region upstream and 1.5-megabase regions downstream of the associated variants.

Thus, we have identified two variants on chromosome 4q25 that are strongly associated with AF in three distinct populations of European descent. The stronger variant also replicates well in a Chinese population in which it is much more common, and it has higher PAR than in populations of European descent. This association is particularly compelling in younger patients and in those with lone AF, but it is also present in older patients with more commonly encountered forms of AF. Although the mechanism for this association is unknown, our results provide a foundation for further studies on the molecular underpinnings of AF.

METHODS SUMMARY

Subjects. The Icelandic cases consisted of all patients diagnosed with AF and/or AFl at the two largest hospitals in the country from 1987 to 2005. The Swedish cases were recruited from 1996 to 2002 as a part of a continuing genetic epidemiology study, the South Stockholm Ischaemic Stroke Study. The cases in the United States were a mixture of stroke patients with a diagnosis of AF and younger consecutive patients with lone AF or AF with a coexisting diagnosis of hypertension. The Hong Kong cases were a collection of stroke and diabetes patients with a diagnosis of AF. The diagnosis of AF was confirmed by a 12-lead electrocardiogram in all study populations.

The Icelandic controls were chosen at random from individuals who had participated in other genetic studies at deCODE, excluding first-degree relatives of patients and controls (Supplementary Table 7). The Swedish controls were recruited from the same region as patients from blood donors (in 2001) and healthy volunteers (1990–94). The US controls were recruited from a large primary care practice and from patients participating in a study of haemorrhagic stroke. The Hong Kong controls were individuals without a diagnosis of AF.

Association analysis. In the genome-wide association stage, Icelandic case and control samples were assayed with Infinium HumanHap300 SNP chips (Illumina), containing 317,511 SNPs, from which 316,515 were polymorphic and satisfied our quality criteria.

A likelihood procedure described previously[19] was used for the association analyses. Allele-specific OR was calculated on the assumption of a multiplicative model[20]. Results from multiple case-control groups were combined by using a Mantel–Haenszel model[21]. In all tables, *P* values for both rs2200733 and rs10033464 were computed on the basis of comparison with the wild-type rs2200733 C, rs13143308 G, rs10033464 G haplotype carrying neither of the at-risk alleles.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 April; accepted 11 June 2007.
Published online 1 July 2007.

1. Go, A. S. *et al.* Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *J. Am. Med. Assoc.* **285,** 2370–2375 (2001).
2. Miyasaka, Y. *et al.* Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. *Circulation* **114,** 119–125 (2006).
3. Arnar, D. O. *et al.* Familial aggregation of atrial fibrillation in Iceland. *Eur. Heart J.* **27,** 708–712 (2006).
4. Fox, C. S. *et al.* Parental atrial fibrillation as a risk factor for atrial fibrillation in offspring. *J. Am. Med. Assoc.* **291,** 2851–2855 (2004).
5. Ellinor, P. T., Yoerger, D. M., Ruskin, J. N. & MacRae, C. A. Familial aggregation in lone atrial fibrillation. *Hum. Genet.* **118,** 179–184 (2005).
6. Chen, Y. H. *et al.* KCNQ1 gain-of-function mutation in familial fibrillation. *Science* **299,** 251–254 (2003).
7. Yang, Y. *et al.* Identification of a KCNE2 gain-of-function mutation in patients with familial atrial fibrillation. *Am. J. Hum. Genet.* **75,** 899–905 (2004).
8. Xia, M. *et al.* A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. *Biochem. Biophys. Res. Commun.* **332,** 1012–1019 (2005).
9. Olson, T. M. *et al.* Kv1.5 channelopathy due to KCNA5 loss-of-function mutation causes human atrial fibrillation. *Hum. Mol. Genet.* **15,** 2185–2191 (2006).
10. Hong, K., Bjerregaard, P., Gussak, I. & Brugada, R. Short QT syndrome and atrial fibrillation caused by mutation in KCNH2. *J. Cardiovasc. Electrophysiol.* **16,** 394–396 (2005).
11. Ellinor, P. T. *et al.* Mutations in the long QT gene, KCNQ1, are an uncommon cause of atrial fibrillation. *Heart* **90,** 1487–1488 (2004).
12. Ellinor, P. T., Petrov-Kondratov, V. I., Zakharova, E., Nam, E. G. & MacRae, C. A. Potassium channel gene mutations rarely cause atrial fibrillation. *BMC Med. Genet.* **7,** 70 (2006).
13. Franco, D. & Campione, M. The role of Pitx2 during cardiac development. Linking left–right signaling and congenital heart diseases. *Trends Cardiovasc. Med.* **13,** 157–163 (2003).
14. Faucourt, M., Houliston, E., Besnardeau, L., Kimelman, D. & Lepage, T. The pitx2 homeobox protein is required early for endoderm formation and nodal signaling. *Dev. Biol.* **229,** 287–306 (2001).
15. Mommersteeg, M. T. *et al.* Molecular pathway for the localized formation of the sinoatrial node. *Circ. Res.* **100,** 354–362 (2007).
16. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
17. Waldo, A. L. The interrelationship between atrial fibrillation and atrial flutter. *Prog. Cardiovasc. Dis.* **48,** 41–56 (2005).

18. Zini, S. *et al.* Identification of metabolic pathways of brain angiotensin II and III using specific aminopeptidase inhibitors: predominant role of angiotensin III in the control of vasopressin release. *Proc. Natl Acad. Sci. USA* **93,** 11968–11973 (1996).
19. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genet.* **35,** 131–138 (2003).
20. Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51,** 227–233 (1987).
21. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22,** 719–748 (1959).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the patients and their family members whose contribution made this work possible; the nurses at Noatun (deCODE's sample recruitment center), personnel at the deCODE core facilities, and M. Shea for the ongoing enrolment of patients at Massachusetts General Hospital; and A. Plourde and S. Makino for technical assistance.

Author Contributions D.F.G., D.O.A., A.H., S.G., P.T.E., J.R. U.T. and K.S. wrote the first draft of the paper. D.O.A., H.H., R.S., J.T.S. and G.T. collected and diagnosed the Icelandic AF samples. Ko.K. and J.H. collected and diagnosed the Swedish samples. K.L.F., S.M.G., M.S., P.K., C.A.M., E.E.S., J.R. and P.T.E. collected and diagnosed the US samples. M.C.Y.N., L.B., W.Y.S., K.S.W., J.C.N.C. and R.C.W.M collected and diagnosed the Hong Kong samples. A.H., S.G., A.S., A.J., A.B., T.B., V.M.B., G.A.H. and E.P. performed genotyping and experimental work. D.F.G., G.T., A.P., P.S., A.H. and A.K. analyzed the data. D.F.G., D.O.A., A.H., S.G., Kr.K., J.R., J.H., R.C.W.M., P.T.E, G.T, J.R.G., A.K., U.T. and K.S. planned, supervised and coordinated the work. All authors contributed to the final version of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.F.G. (daniel.gudbjartsson@decode.is) or K.S. (kstefans@decode.is).

METHODS

Icelandic study population. This study initially included all the patients consenting to participation who were diagnosed with AF and/or AFl (ICD (International Classification of Diseases) 10 diagnosis I48 and ICD 9 diagnosis 427.3) at Landspitali University Hospital in Reykjavik, the only tertiary referral centre in Iceland, and at Akureyri Regional Hospital, the second largest hospital in the country, from 1987 to 2005. All diagnoses were confirmed by a 12-lead electrocardiogram (ECG), which was read manually by a cardiologist. All cases were included, regardless of whether or not the patients had clinical symptoms, except those diagnosed only immediately after open cardiac surgery.

A set of 550 cases were successfully genotyped in accordance with our quality control criteria in a genome-wide SNP genotyping effort, using the Infinium II assay method and the Sentrix HumanHap300 BeadChip (Illumina). The age at diagnosis for this initial group of 550 patients (370 males and 180 females) was 72.5 ± 11.0 years (mean ± s.d.) and the range was 34.7–96.2 years. The validation group of 2,273 patients (1,359 males and 913 females) had an age at diagnosis of 70.5 ± 13.0 years and the range was 16.8–100.6. The AF/AFl-free controls (2,201 males and 2,275 females at the initial genome-wide screening aged 61.5 ± 15.8 years and 5,654 males and 7,597 females at the validation stage aged 61.9 ± 18.4 years) used in this study consisted of controls randomly selected from the Icelandic genealogical database and individuals from other ongoing related genetic studies at deCODE. Controls with first-degree relatives (siblings, parents or offspring) with AF/AFl, or a first-degree control relative, were excluded from the analysis.

The study was approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. Written informed consent was obtained from all patients, relatives and controls. Personal identifiers associated with medical information and blood samples were encrypted with a third-party encryption system as described previously[22].

Swedish study population. Patients with ischaemic stroke or transient ischaemic attack attending the stroke unit or the stroke outpatient clinic at Karolinska University Hospital, Huddinge unit, in Stockholm, Sweden, were recruited from 1996 to 2002 as part of an ongoing genetic epidemiology study, the South Stockholm Ischaemic Stroke Study (SSISS). The study was approved by the Bioethics Committee of Karolinska Institutet (Dnr 286/96 and 08/02). A diagnosis of AF in the Swedish samples was based on a 12-lead ECG. The fraction of males in the Swedish AF cases was 46.2% and the age at stroke diagnosis for the Swedish AF cases was 74.4 ± 8.7 years.

The Swedish controls used in this study were population-based controls recruited from the same region in central Sweden as the patients, representing the general population in this area. The individuals were either blood donors (recruited in 2001) or healthy volunteers (collected in 1990–94) recruited by the clinical chemistry department at the Karolinska University Hospital to represent a normal reference population. The fraction of males in the Swedish controls was 59.7% and the age at recruitment for the Swedish controls was 43.1 ± 12.3 years.

Study population in the United States. US subjects were enrolled in ongoing case-control and cohort studies at Massachusetts General Hospital between January 1998 and July 2006. All aspects of these studies have been approved by the local Institutional Review Board. Subjects who were enrolled in the case-control study consisted of patients hospitalized with acute ischaemic or haemorrhagic stroke confirmed by computed tomography or magnetic resonance imaging, admitted to a single acute-care hospital. Of the 328 haemorrhagic stroke patients recruited, 78 were diagnosed with AF and were used as cases for the current study; the remaining 250 were used as controls. A total of 170 ischaemic stroke patients had a diagnosis of AF and were treated as cases, but no ischaemic stroke patients were treated as controls. Patients were excluded for primary subarachnoid haemorrhage and for intracerebral haemorrhage secondary to head trauma, tumour, vascular malformation, or vasculitis. A total of 624 stroke-free controls were recruited from a large, primary care practice (more than 18,000 patients) serving the hospital's catchment area as well as the hospital's Anticoagulation Management Service; 70 of the 624 individuals collected as controls were diagnosed with AF and treated as cases for the purposes of the current study. Of all individuals used as controls, 50.9% were males and their age was 67.4 ± 12.3 years. All subjects or an accompanying informant provided informed consent for participation in genetic studies and were interviewed prospectively about medical history, medications and social and family history. The presence or absence of AF was documented prospectively through interviews and from a review of medical records.

The second part of the US subjects consisted of consecutive patients with lone AF or AF with a coexisting diagnosis of hypertension referred to the arrhythmia service who provided written informed consent for participation in genetic studies. Inclusion criteria were AF documented by ECG, and an age of 65 years or less. The exclusion criteria were structural heart disease as assessed by echocardiography, rheumatic heart disease, hyperthyroidism, myocardial infarction or congestive heart failure. Each patient underwent a physical examination and a standardized interview to identify past medical conditions, medications, symptoms and possible triggers for the initiation of AF. All patients were evaluated by 12-lead ECG, echocardiogram and laboratory studies. ECGs and echocardiograms were interpreted by using standard criteria.

Study population in Hong Kong. All subjects in the Hong Kong study population were of southern Han Chinese ancestry residing in Hong Kong. The cases consisted of 217 individuals (49.1% male, aged 68.1 ± 9.6 years) selected from the Prince of Wales Hospital Diabetes Registry[23], and 116 subjects (30.2% male, aged 76.1 ± 10.9 years) from the Stroke Registry[24]. All subjects were diagnosed by ECG as having AF. The controls consisted of 2,836 subjects without evidence of AF. Informed consent was obtained for each participating subject. This study was approved by the Clinical Research Ethics Committee of the Chinese University of Hong Kong.

22. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genet.* **38,** 320–323 (2006).
23. Yang, X. *et al.* Development and validation of stroke risk equation for Hong Kong Chinese patients with type 2 diabetes: the Hong Kong Diabetes Registry. *Diabetes Care* **30,** 65–70 (2007).
24. Baum, L. *et al.* Methylenetetrahydrofolate reductase gene A222V polymorphism and risk of ischemic stroke. *Clin. Chem. Lab. Med.* **42,** 1370–1376 (2004).

LETTERS

Rhythmic growth explained by coincidence between internal and external cues

Kazunari Nozue[1], Michael F. Covington[1], Paula D. Duek[2]†, Séverine Lorrain[2], Christian Fankhauser[2], Stacey L. Harmer[1] & Julin N. Maloof[1]

Most organisms use circadian oscillators to coordinate physiological and developmental processes such as growth with predictable daily environmental changes like sunrise and sunset. The importance of such coordination is highlighted by studies showing that circadian dysfunction causes reduced fitness in bacteria[1] and plants[2], as well as sleep and psychological disorders in humans[3]. Plant cell growth requires energy and water—factors that oscillate owing to diurnal environmental changes. Indeed, two important factors controlling stem growth are the internal circadian oscillator[4–6] and external light levels[7]. However, most circadian studies have been performed in constant conditions, precluding mechanistic study of interactions between the clock and diurnal variation in the environment. Studies of stem elongation in diurnal conditions have revealed complex growth patterns, but no mechanism has been described[8–10]. Here we show that the growth phase of *Arabidopsis* seedlings in diurnal light conditions is shifted 8–12 h relative to plants in continuous light, and we describe a mechanism underlying this environmental response. We find that the clock regulates transcript levels of two basic helix–loop–helix genes, phytochrome-interacting factor 4 (*PIF4*) and *PIF5*, whereas light regulates their protein abundance. These genes function as positive growth regulators; the coincidence of high transcript levels (by the clock) and protein accumulation (in the dark) allows them to promote plant growth at the end of the night. Thus, these two genes integrate clock and light signalling, and their coordinated regulation explains the observed diurnal growth rhythms. This interaction may serve as a paradigm for understanding how endogenous and environmental signals cooperate to control other processes.

Most core circadian clocks consist of a conserved oscillatory mechanism using a transcriptional negative feedback loop. In plants, two Myb-like transcription factors (CIRCADIAN CLOCK-ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCOTYL (LHY)) and a pseudo-response regulator (TIMING OF CAB EXPRESSION 1 (TOC1)) are thought to be components of the classical central oscillator[11]. Overexpression of *CCA1* or *LHY* causes arrhythmia in circadian-controlled gene expression and growth[6,12,13]. In addition to the core components, *EARLY FLOWERING 3* (*ELF3*) is required to restrict light input to the clock and other signalling pathways[14,15]. Under constant dim light, hypocotyl elongation is rhythmically controlled in a process that requires *CCA1* and ELF3 (refs 6, 16).

Light, perceived by phytochrome and cryptochrome photoreceptors, strongly reduces seedling growth rate[7]. Phytochromes signal in part by inducing degradation of a family of basic helix–loop–helix transcription factors known as PIFs or PIF3-likes (PILs) that, when present, act primarily to inhibit light responses[17]. The phytochrome and cryptochrome signalling pathways converge at other transcription factors, including the basic leucine zipper factor HY5 (ref. 18).

To examine how internal rhythms and photoperception interact to control growth, we asked whether rhythmic hypocotyl growth is light-dependent. As previously reported[6], under continuous light wild-type hypocotyls exhibited rhythmic growth, peaking at subjective dusk, whereas the arrhythmic mutant *elf3* grew continuously (Fig. 1a and Supplementary Fig. 1). Although the clock is known to function in continuous darkness[3], growth in continuous darkness was rapid and arrhythmic (Fig. 1b). This indicated that observable circadian growth control is light-dependent, and suggested that the clock and photoreceptor signalling pathways might interact to control normal growth. To investigate the interactions further, plants growing in short-day conditions were examined. The growth pattern of wild-type hypocotyls under short-day conditions was strikingly different from that seen in continuous light: peak growth occurred at dawn instead of at subjective dusk (Fig. 1c). This interaction between the clock and light signalling has been missed in previous studies because they were performed in continuous light conditions.

To investigate further the requirement for light, we asked whether photoreceptor signalling is required for rhythmic growth. Growth rhythms were weak or absent in *hy5* mutants (Fig. 1c), which are deficient in both phytochrome and cryptochrome signalling[18]. Additionally, rhythmic growth was very weak or absent in *hy2* mutants, which lack the phytochrome chromophore[19], suggesting that phytochrome signalling is necessary for rhythmic growth. The observed growth patterns could reflect the kinetics of signal transduction through the photomorphogenic pathways directly, or could result from interactions between the clock and light signalling. To distinguish these possibilities, we examined growth patterns in arrhythmic clock mutants. Unlike the wild type, arrhythmic *CCA1*-overexpressing (*CCA1*-OX) and *elf3* plants seemed to respond directly to changes in light: growth ceased at dawn and resumed at dusk (Fig. 1c). Quantification of these results using a dark responsiveness index (see Methods) confirmed that arrhythmic *elf3* and *CCA1*-OX plants were considerably more responsive to darkness than the wild type or light-signalling mutants (Supplementary Fig. 2). These results show: first, that arrhythmic and light-signalling mutants have distinct growth patterns, suggesting that they do not impair the same pathway; second, that the transition from light to darkness can quickly increase elongation rates, but these effects are normally buffered by the clock; and, third, that under diurnal cycles the clock functions to maintain the growth-repressing function of light during the first half of the night. (See Supplementary Discussion and Supplementary Figs 3, 4, 7 and 8 for detailed discussion of growth patterns in other clock mutants and diurnal versus continuous photoperiods.)

To confirm this hypothesis, we examined growth under an 8 h T-cycle (repeating 4 h light:4 h dark treatments, 4L:4D; Fig. 1d). Wild-type plants exhibited frequency demultiplication[3]: growth cues from

[1]Section of Plant Biology, College of Biological Sciences, University of California, Davis, One Shields Avenue, Davis, California 95616, USA. [2]Center for Integrative Genomics, University of Lausanne, Genopode Building, CH-1015 Lausanne, Switzerland. †Present address: Swiss Institute of Bioinformatics, 1 rue Michel Servet, CH-1211 Geneva 4, Switzerland.

two 4L:4D cycles among three were ignored, producing a 24 h rather than an 8 h rhythm. A simple interpretation is that the circadian clock gates growth in the dark. However, because rhythmic growth is not seen in entrained plants shifted to constant darkness (Fig. 1b), previous light exposure is required to see this effect. Thus, the clock gates the rate at which light-signalling is inactivated on transfer to darkness. Under these 8 h T-cycles, *CCA1*-OX and *elf3* grew during every dark period (Fig. 1d), confirming that plants without a functional clock respond directly to light/dark transitions and that the clock is responsible for the frequency demultiplication seen in the wild type.

Clock-regulated transcription is thought to be an important mechanism for circadian control of physiological processes[3]. We therefore hypothesized that the clock might regulate hypocotyl elongation at the transcriptional level. To find genes that might link the clock and light-signalling pathways, we compared expression profiles in plants that either did or did not respond to darkness with growth. Specifically, we sampled both wild-type Columbia (Col) and *CCA1*-OX plants at two time points in the dark, one when Col is growing and one when it is not (see Supplementary Fig. 5 and Supplementary Discussion). When we used whole-genome expression analysis to contrast both elongating versus non-elongating Col and elongating *CCA1*-OX versus non-elongating Col, we identified 38 genes that are upregulated and 23 genes that are downregulated in growing plants (Supplementary Tables 1, 2 and Supplementary Discussion). Among these, we identified two closely related basic helix–loop–helix genes, *PIF4* (ref. 20) and *PIF5* (also known as *PIL6*; refs 21, 22), as the two strongest candidates on the basis of their high ranking by rank-product analysis (Fig. 2a and Supplementary Table 1 and previously published data). These two genes code for phytochrome-B-interacting proteins that negatively regulate light signalling and positively regulate growth[20,22,23]. Expression of *PIF4* and *PIF5* messenger RNA is clock-regulated in continuous light[21] (Supplementary Fig. 6), and both proteins interact with a central clock component, TOC1 (ref. 17), suggesting *PIF4* and *PIF5* as possible links between the clock and light signalling. Because knockout or overexpression of either gene does not alter central clock properties (Supplementary Fig. 6), there seems to be a regulatory cascade: clock to PIF4 and PIF5, and from PIF5 to light signalling, as previously suggested for PIF5 (ref. 23).

To complement our microarray data (Fig. 2a), we examined expression of *PIF4* and *PIF5* in plants grown in short-day conditions by quantitative PCR with reverse transcription (qRT–PCR) (Fig. 2b, c). In wild-type plants, expression of both genes decreased soon after lights were turned off and began to rise in the mid- to late-night, correlating well with the observed phase of hypocotyl growth in short-day conditions. In *CCA1*-OX plants, these genes cycle with low amplitude, if at all, such that levels are much higher during the early and middle night in *CCA1*-OX compared with the wild type, which is consistent with growth patterns in this genotype.

We hypothesized that lack of plant growth in the early night might be due to clock-mediated repression of *PIF4* and *PIF5* expression. To test this, we examined growth in plants overexpressing *PIF4* or *PIF5*. Similar to the arrhythmic lines, these plants grew immediately in response to darkness both in short-day conditions and in 8 h T-cycles (Fig. 2d, e, and Supplementary Fig. 2). These plants do not show decreased growth in the light during the first few days of the short-day experiment, probably owing to the very high levels of *PIF4* or *PIF5* expression—approximately ten times higher than in *CCA1*-OX (Supplementary Fig. 6). Together with the microarray and qRT–PCR expression data, the growth patterns in the OX lines show that the clock gates the dark growth response by regulating *PIF4* and *PIF5* expression.

Hypocotyl growth is inhibited by light in wild-type and *PIF4*- and *PIF5*-OX plants, as well as in *CCA1*-OX plants, indicating that *PIF4* and *PIF5* mRNA levels are not the only factors determining growth. Light triggers degradation of PIF3 and PIF1 (also known as PIL5) proteins[24–26]—close paralogues of PIF4 and PIF5. Indeed, we found that PIF4 and PIF5 protein levels rapidly decrease in the light and increase in response to darkness (Fig. 2f and g), correlating well with observed growth responses in overexpressors (Fig. 2d and e). This correlation strongly supports the idea that PIF4 and PIF5 proteins are key regulators of diurnal growth rhythms. This notion is further supported by *pif4* or *pif5* single-mutant or *pif4 pif5* double-mutant plants, which show a partial or complete loss, respectively, of the dawn growth peak (Fig. 2e). Thus, PIF4 and PIF5 are partially redundant in function, but both are required to fully promote growth in diurnal light conditions.

On the basis of our observations we propose a model for circadian and light regulation of growth (Fig. 2h): during the day, light inhibits

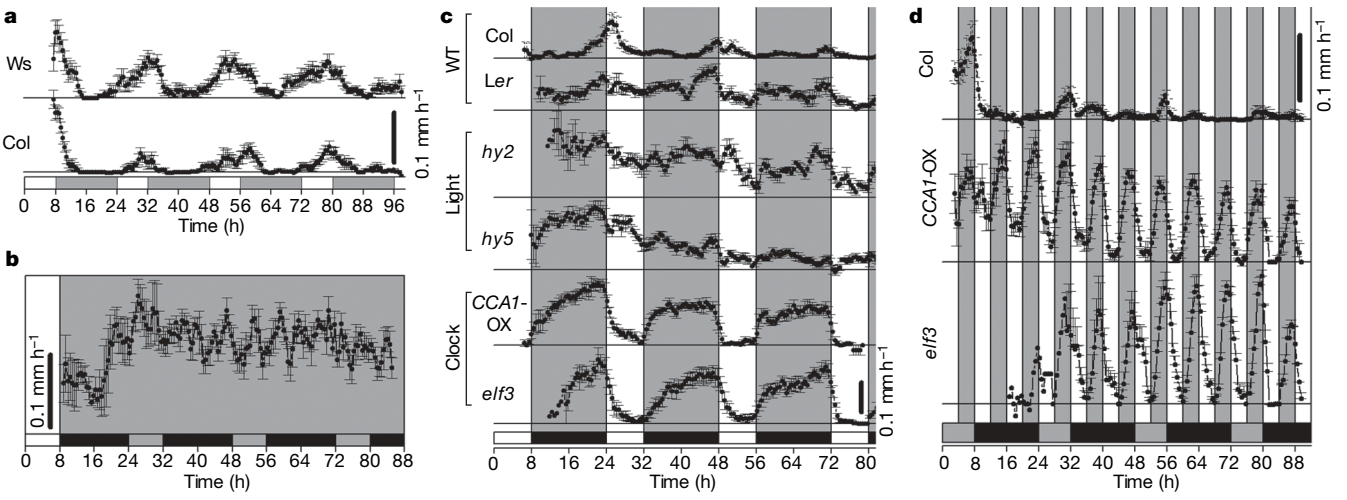


Figure 1 | Diurnal rhythms of hypocotyl elongation require light and the circadian clock. Plants were entrained for three days under short-day conditions and then switched to continuous light (**a**) or 4L:4D (**d**). Alternatively, plants were entrained for four days under short-day conditions and then switched to continuous darkness (**b**) or kept in short-day conditions (**c**). We used infrared imaging to monitor seedling growth (see Methods). Growth rate is plotted as a function of time; zero indicates dawn of the fourth day. The vertical scale bar indicates 0.1 mm h^{-1}. Measurements were started when hypocotyls were easily discernible, typically $t = 8$. The mean ± s.e.m. of at least two independent experiments is shown; $n \geq 6$ seedlings. In all plot areas, times of true light and darkness are indicated by clear and grey rectangles, respectively; see below for meaning of *x* axis rectangles. **a**, Rhythmic elongation of wild-type Col and Wassilewskija (Ws) hypocotyls in continuous light. White and grey bars on the *x* axis indicate subjective day and night, respectively. **b**, Continuous hypocotyl elongation of wild-type (WT) Col in continuous darkness. Grey and black bars on the *x* axis indicate subjective day and night, respectively. **c**, Hypocotyl elongation in short-day conditions. Col is the wild-type background for *CCA1*-OX, *elf3* and *hy2*; Landsberg *erecta* (L*er*) is the wild-type background for *hy5*. **d**, Growth in 4L:4D conditions is altered in clock mutants. Grey and black rectangles on the *x* axis indicate subjective day and night, respectively.

growth in part by inactivating the growth-promoting transcription factors PIF4 and PIF5, probably by degradation (S.L., P.D.D. and C.F., unpublished). During the first half of the night, the clock prevents growth by repressing transcription of *PIF4* and *PIF5*, maintaining the light-signalling pathway in an activated state. Closer to dawn, clock-mediated transcriptional repression of PIF4 and PIF5 fades, allowing their expression and the subsequent promotion of growth. Light is necessary for growth rhythms because it is required to inactivate or degrade PIF4 and PIF5; similarly, normal growth rhythms depend on the downregulation of the *PIF4* and *PIF5* message by the clock. Thus, normal diurnal growth patterns depend on interactions between internal (circadian) and external (light) cues. This is an example of an external coincidence model, originally proposed by Bünning[27] to explain photoperiodic regulation of a seasonal response—the transition from vegetative to reproductive growth.

It has been suggested that strong responses to light/dark transitions obscure circadian regulation of growth[6]. However, our results show that the opposite is true. The clock is critical in specifying wild-type growth patterns under light/dark cycles and can block acute responses to light/dark transitions. Circadian gating of acute responses is widely

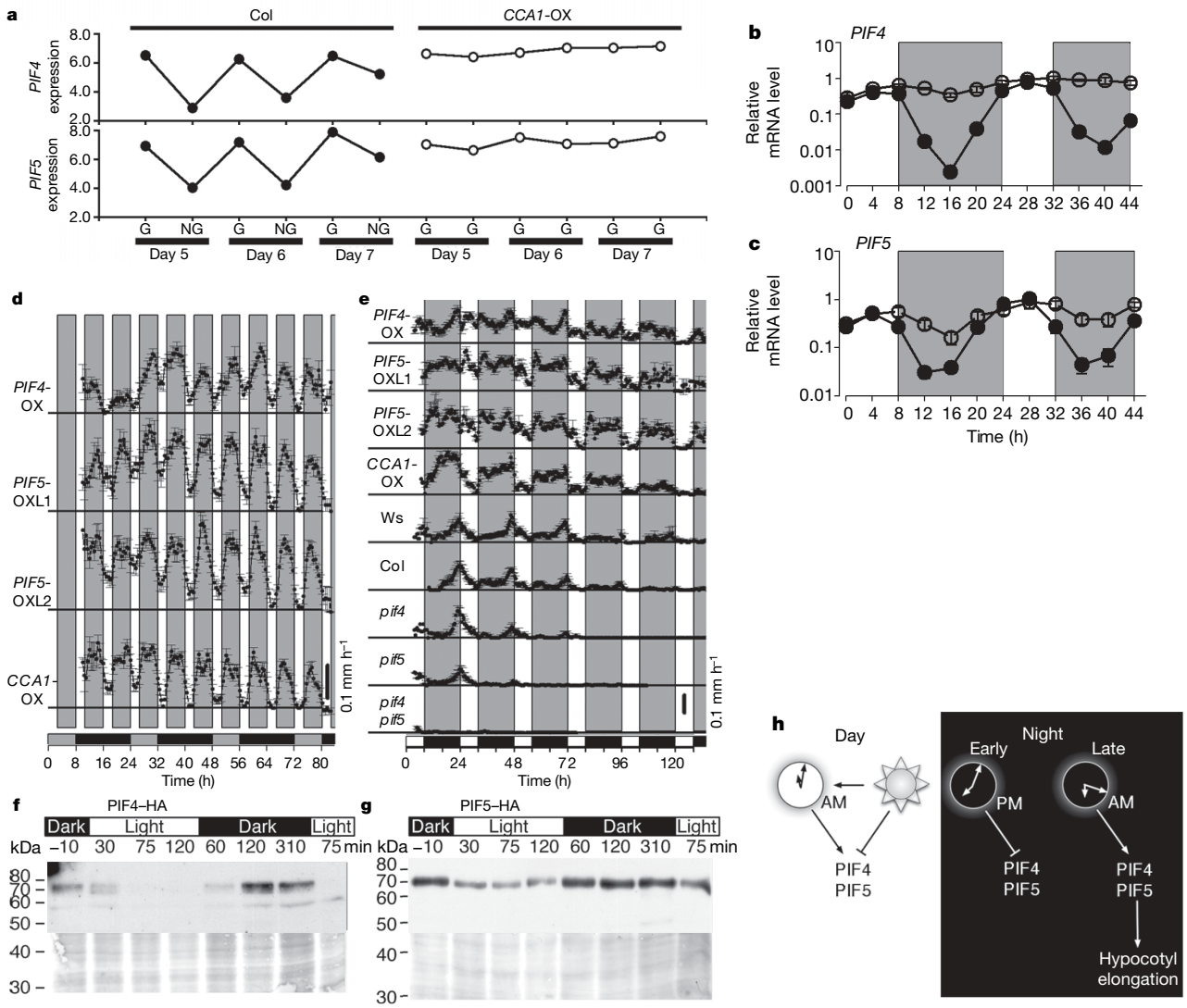


Figure 2 | Transcript and protein level regulation of light-signalling components in hypocotyl growth control. **a**, *PIF4* and *PIF5* expression patterns as assayed by microarray during 160 min light:320 min dark cycles (the light and dark period of a short day divided by 3, SD/3) are correlated with timing of hypocotyl elongation. Letters on the *x* axis indicate observed growth (G) or non-growth (NG) phenotypes at the time of collection. **b**, **c**, Diurnal expression patterns of *PIF4* (**b**) and *PIF5* (**c**) mRNA in wild-type and *CCA1*-OX plants as determined by qRT–PCR. Filled and open circles indicate expression level in Col and *CCA1*-OX, respectively. Zero indicates dawn of the fifth day. Shaded rectangles indicate darkness. The mean ± s.e.m. of two independent experiments, each with two replicates, are shown. **d**, *PIF4* and *PIF5* overexpression or knockout impairs clock regulation of growth. The experiment was performed as in Fig. 1d except SD/3 cycles were used instead of 4L:4D cycles and entrainment occurred over four days instead of three days. OXL1 and OXL2 refer to two independent lines overexpressing *PIF5* (see Methods). **e**, *PIF4* and *PIF5* overexpression or knockout alter rhythmic diurnal growth patterns. The experiment was performed as in Fig. 1c. **f**, **g**, PIF4 and PIF5 protein levels decrease in the light and increase in the dark. Lines constitutively expressing HA-tagged PIF4 (**f**) and PIF5 (**g**) were grown under SD/3. Times (in minutes) after transitions from dark to light and from light to dark are indicated. HA-tagged protein levels during the first cycle of day seven are shown in upper panels. Protein loading is shown in lower panels by Coomassie staining of blotted membranes. Two independent experiments were performed for two independent transgenic lines, and representative figures are shown here (PIF4–HA;OX5 in **f** and PIF5–HA;OX3 in **g**). kDa, kilodalton. **d**, **e**, *PIF4*-OX is driven by the PIF4 promoter, but expression is 25-fold higher than in the wild type[22]; **f**, *PIF4–HA* is driven from the cauliflower mosaic virus 35S promoter. **g**, Overexpression of PIF5 and PIF5–HA was achieved using the cauliflower mosaic virus 35S promoter. **h**, External coincidence model for rhythmic growth generation; see text for details.

known[3], but the mechanisms are mostly unidentified. Our finding that a dark-induced growth response is transcriptionally regulated by the clock provides a model that can be used to examine other gated responses. In addition, our finding that growth-promoting transcription factors are controlled by light and the clock by means of different regulatory mechanisms could provide a model for how other types of signalling pathways converge on common regulatory factors.

It is worth considering an ecological reason for rapid growth during late night. One possibility is that this allows plants to time growth to coincide with maximum water availability, because growth is among the first responses to be limited by water[28,29]. Another possibility is that this system allows plants to buffer their responses to acute changes in light, thereby growing only in response to extended periods of darkness.

Many modern studies of biological rhythms have been carried out in constant environmental conditions—a reductionist approach that has yielded important insights into the nature of the circadian clock. However, an understanding of how the circadian system functions in the real world will require more complex experimental conditions that better approximate the natural world. Using such conditions, we have uncovered a novel interaction among well-known regulatory networks that together regulate plant growth. Further investigation of how molecular networks respond to diurnal conditions will improve our understanding of how organisms live in and respond to their natural environment.

METHODS SUMMARY

Plant materials and growth condition. Plant materials are described in detail in the Methods. Sterilized seeds were plated in square plates. Seeds were stratified for four days and incubated under short-day cycles (8 h white fluorescent light:16 h dark (short-day conditions)) for three days. In most cases, monitoring of plant growth started from day four.

Time-lapse photography and image analysis. To acquire images in the dark, five seconds of background infrared illumination was given by infrared light-emitting diode (infrared-LED), and plant images were captured by an infrared-sensitive charge-coupled device camera. The infrared-LED was controlled via a USB port. Images were captured at 30-min intervals. Hypocotyl length was measured by Image J software. Data analysis was performed in the statistical environment R (http://www.R-project.org/).

RNA extraction, whole-genome expression analysis and qRT–PCR analysis. RNA was extracted by RNeasy Plant Mini kit (Qiagen). Affymetrix ATH1 genome arrays were used for the whole-genome expression assay. Robust multi-array averaging and rank product analysis were performed using Bioconductor in the R environment (Bioconductor /R) Detailed information is found in the Methods.

Western blot analysis. Details of generation of *Arabidopsis* plants overexpressing haemagglutinin (HA)-tagged PIF4 or PIF5 protein and methods for detecting PIF4–HA or PIF5–HA proteins levels are described in the Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 January; accepted 16 May 2007.
Published online 24 June 2007.

1. Woelfle, M. A., Ouyang, Y., Phanvijhitsiri, K. & Johnson, C. H. The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. *Curr. Biol.* **14,** 1481–1486 (2004).
2. Dodd, A. N. *et al.* Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* **309,** 630–633 (2005).
3. Dunlap, J. C., Loros, J. J. & DeCoursey, P. J. (eds) *Chronobiology* (Sinauer Associates, Sunderland, Massachusetts, 2005).
4. Lecharny, A. & Wagner, E. Stem extension rate in light-grown plants. Evidence for an endogenous circadian rhythm in *Chenopodium. Physiol. Plant.* **60,** 437–443 (1984).
5. Ibrahim, C. A., Lecharny, A. & Millet, B. Circadian endogenous growth rhythm in tomato. *Plant Physiol.* **67,** 113 (1981).
6. Dowson-Day, M. J. & Millar, A. J. Circadian dysfunction causes aberrant hypocotyl elongation patterns in *Arabidopsis. Plant J.* **17,** 63–71 (1999).
7. Chen, M., Chory, J. & Fankhauser, C. Light signal transduction in higher plants. *Annu. Rev. Genet.* **38,** 87–117 (2004).
8. Bertram, L. & Karlsen, P. Patterns in stem elongation rate in chrysanthemum and tomato plants in relation to irradiance and day/night temperature. *Sci. Hortic.* **58,** 139–150 (1994).
9. Tutty, J. R., Hicklenton, P. R., Kristie, D. N. & McRae, K. B. The influence of photoperiod and temperature on the kinetics of stem elongation in *Dendranthema grandiflorum. J. Am. Soc. Hortic. Sci.* **119,** 138–143 (1994).
10. Bertram, L. & Lercari, B. Kinetics of stem elongation in light-grown tomato plants. Responses to different photosynthetically active radiation levels by wild-type and *aurea* mutant plants. *Photochem. Photobiol.* **66,** 396–403 (1997).
11. Gardner, M. J., Hubbard, K. E., Hotta, C. T., Dodd, A. N. & Webb, A. A. How plants tell the time. *Biochem. J.* **397,** 15–24 (2006).
12. Schaffer, R. *et al.* The late elongated hypocotyl mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. *Cell* **93,** 1219–1229 (1998).
13. Wang, Z. Y. & Tobin, E. M. Constitutive expression of the *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) gene disrupts circadian rhythms and suppresses its own expression. *Cell* **93,** 1207–1217 (1998).
14. McWatters, H. G., Bastow, R. M., Hall, A. & Millar, A. J. The ELF3 *zeitnehmer* regulates light signalling to the circadian clock. *Nature* **408,** 716–720 (2000).
15. Covington, M. F. *et al. ELF3* modulates resetting of the circadian clock in *Arabidopsis. Plant Cell* **13,** 1305–1315 (2001).
16. Thain, S. C. *et al.* Circadian rhythms of ethylene emission in *Arabidopsis*. *Plant Physiol.* **136,** 3751–3761 (2004).
17. Duek, P. D. & Fankhauser, C. bHLH class transcription factors take centre stage in phytochrome signalling. *Trends Plant Sci.* **10,** 51–54 (2005).
18. Osterlund, M. T., Hardtke, C. S., Wei, N. & Deng, X. W. Targeted destabilization of HY5 during light-regulated development of *Arabidopsis. Nature* **405,** 462–466 (2000).
19. Kohchi, T. *et al.* The *Arabidopsis HY2* gene encodes phytochromobilin synthase, a ferredoxin-dependent biliverdin reductase. *Plant Cell* **13,** 425–436 (2001).
20. Huq, E. & Quail, P. H. PIF4, a phytochrome-interacting bHLH factor, functions as a negative regulator of phytochrome B signaling in *Arabidopsis*. *EMBO J.* **21,** 2441–2450 (2002).
21. Yamashino, T. *et al.* A link between circadian-controlled bHLH factors and the *APRR1/TOC1* quintet in *Arabidopsis thaliana*. *Plant Cell Physiol.* **44,** 619–629 (2003).
22. Khanna, R. *et al.* A novel molecular recognition motif necessary for targeting photoactivated phytochrome signaling to specific basic helix–loop–helix transcription factors. *Plant Cell* **16,** 3033–3044 (2004).
23. Fujimori, T., Yamashino, T., Kato, T. & Mizuno, T. Circadian-controlled basic/helix–loop–helix factor, *PIL6*, implicated in light-signal transduction in *Arabidopsis thaliana. Plant Cell Physiol.* **45,** 1078–1086 (2004).
24. Park, E. *et al.* Degradation of phytochrome interacting factor 3 in phytochrome-mediated light signaling. *Plant Cell Physiol.* **45,** 968–975 (2004).
25. Shen, H., Moon, J. & Huq, E. PIF1 is regulated by light-mediated degradation through the ubiquitin–26S proteasome pathway to optimize photomorphogenesis of seedlings in *Arabidopsis*. *Plant J.* **44,** 1023–1035 (2005).
26. Al-Sady, B., Ni, W., Kircher, S., Schafer, E. & Quail, P. H. Photoactivated phytochrome induces rapid PIF3 phosphorylation prior to proteasome-mediated degradation. *Mol. Cell* **23,** 439–446 (2006).
27. Bünning, E. Die endogene Tagesrhythmik als Grundlage der photoperiodischen reaktion. *Ber. Dtsch. Bot. Ges.* **54,** 590–607 (1936).
28. Tsuda, M. & Tyree, M. T. Plant hydraulic conductance measured by the high pressure flow meter in crop plants. *J. Exp. Bot.* **51,** 823–828 (2000).
29. Hsiao, T. C. Plant responses to water stress. *Annu. Rev. Plant Physiol.* **24,** 519–570 (1973).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Wallace for technical assistance; J. C. Lagarias, N. Sinha and C. Wessinger for critical reading and comments on the manuscript; E. Tobin, S. Kay, A. Millar, J. C. Lagarias, T. Mizuno, P. Quail and the *Arabidopsis* Biological Resources Centre for seeds; and J. C. Lagarias for the loan of computer equipment. This work was supported by grants from the NSF (to J.N.M. and C. Weinig), the Swiss National Science foundation (to C.F.), the HFSP (to C.F., J.N.M. and U. Genick), the NRI of the USDA CSREES (to M.F.C.) and the NIH (to S.L.H.).

Author Contributions K.N. performed all experiments. Statistical analysis of growth and microarray data was done by J.N.M. K.N. and J.N.M. wrote the paper. P.D.D., S.L. and C.F. contributed HA-tagged protein overexpressing plants, western blot protocols, and *pif4 pif5* double-mutant seed. M.F.C. contributed microarray experimental design. K.N., J.N.M. and S.L.H. contributed to project design. All authors discussed the results and commented on the manuscript.

Author Information The microarray data have been deposited in the GEO database (http://www.ncbi.nlm.nih.gov/projects/geo/) under accession number GSE6906. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.N.M. (jnmaloof@ucdavis.edu).

METHODS

Plant materials and growth conditions. Seeds of *CCA1*-OX (also known as *CCA1-34*) (ref. 13), *toc1-2* (ref. 30), *elf4-1* (ref. 31) and *hy2-103* (ref. 19) were provided by E. Tobin (University of California), S. Kay (The Scripps Research Institute), A. Millar (University of Edinburgh) and J. C. Lagarias (University of California, Davis), respectively. 35S-driven *PIF5*-OX lines (PIF5-OXL1 and 2, used in Fig. 2d and e and in Supplementary Figs 2 and 6) and *pif5*-knockout seeds[23] were provided by T. Mizuno (University of Nagoya). *PIF4*-OX and *pif4*-knockout (also called *slr2*) seeds[20] were provided by P. Quail (University of California, Berkeley). In this PIF4 overexpression line, designated *PIF4-OX* in this paper, PIF4 is driven by its native promoter but is expressed approximately 25-fold higher than in the wild type, presumably owing to the insertion site of the transgene[22]. This line was used for Fig. 2d and e and Supplementary Figs 2 and 6. Wild-type seeds (Col, C24, L*er* and Ws), *elf3-1*, *gi-2* and *hy5-1* were supplied by the *Arabidopsis* Biological Resources Center. Isolation of *pif4-101*, a T-DNA insertion allele, *pif4-101 pif5* double-mutants and plants overexpressing PIF4–HA or PIF5–HA constructs will be described elsewhere (S.L., P.D.D. and C.F., unpublished). The *PIF4–HA* and *PIF5–HA* genes are driven from the 35S promoter and were used for experiments shown in Fig. 2f and g. These constructs rescue their respective knockouts, showing that they encode functional proteins (data not shown). *PIF5–HA* shows growth patterns similar to the untagged *PIF5*-OX, whereas the *PIF4–HA* growth pattern is more similar to the wild type, owing to low expression relative to *PIF4*-OX.

For hypocotyl growth rate measurements, seeds were surface-sterilized with 70% ethanol and 0.1% Triton X-100 for 5 min followed by 95% ethanol for 1 min. Sterilized seeds were resuspended in sterile water and plated in two rows on 40 ml of 1 × MSMO (Murashige and Skoog minimal organics medium without sucrose; M6899, Sigma), 3% sucrose and 1% Phytagar (Invitrogen) in a 10 cm square plate (4021, Nalge Nunc). Seeds were stratified at 4 °C for four days and incubated under short-day cycles for three days. In most cases, monitoring of plant growth started from day four. Light was provided by cool white fluorescent lamps (OSRAM Sylvania) with a fluence rate of 62 μmol m^{-2} s^{-1}. Three layers of screens (Silver Gray Fibreglass, Phifer Wire Products) were used to reduce light intensity to 12 μmol m^{-2} s^{-1} for continuous light and long-day experiments; one layer of screen and adjustment of distance from the light source was used to reduce light intensity to 47 μmol m^{-2} s^{-1} for 12L:12D experiments. The Petri dishes were tilted about 30° from the vertical during entrainment and assays.

For RNA extractions for microarray analysis, ~100 seedlings were grown on filter paper laid on 1 × MSMO, 3% sucrose and 0.9% agar (Cat A1296, Sigma) in 6 cm round Petri dishes. Seeds were stratified as above and then incubated for four days at 20 °C under short-day conditions. The Petri dishes were tilted about 30° from vertical in a growth chamber equipped with red and blue LEDs (70.7 μE red light and 15.3 μE blue light). Starting from subjective dawn at day five, cycles of 160 min light and 320 min dark (SD/3) were given for three days (Supplementary Fig. 5). Samples were collected two hours after lights were turned off, transferred to 1.5 ml tubes, and immediately frozen by liquid nitrogen and stored at −80 °C. Samples were collected using infrared goggles (Night Vision model NCB4, Night Owl Optics) to avoid exposing plants to visible light.

For qRT–PCR, 10–20 seedlings were grown on 1 × MSMO, 3% sucrose and 0.9% agar plates without filter paper. Seedlings were collected at four-hour intervals from day four to day six (for those under short-day conditions) or from day five to day six under continuous light after four days of short-day entrainment (for those in continuous light).

For western blot analysis, 10–20 seedlings were grown on 1 × MSMO, 3% sucrose and 0.9% agar plates without filter paper. Plants were entrained in short-day conditions for four days and then switched to SD/3 conditions. During the first SD/3 cycle of the ninth day, seedlings were collected 10 min before lights were turned on, then after 30, 75 or 120 min in the light, then after 60, 120 or 310 min in the subsequent dark period, and finally after 75 min in the following light period.

Time-lapse photography and image analysis. To acquire images in the dark, five seconds of background infrared illumination was given by infrared-LED (276-143, Radioshack) with a diffusion filter (R111, Rosco), and plant images were captured by an infrared-sensitive charge-coupled device camera (MEGA-DCS, Videre Design) equipped with a close-up lens (HF25HA-1B, Fujinon). The infrared-LED was controlled by a USB controller (RUSB-PO8/8(R), Rabbit House) using a C^{++} program. Images (1,280 × 960 pixels) were captured at 30-min intervals by a custom-written C^{++} program through a FireWire connection. Hypocotyl length was measured by Image J software (http://rsb.info.nih.gov./ij/). Data analysis was performed in the statistical environment R (ref. 32; http://www.R-project.org/). To compare dark responsiveness of different genotypes (Supplementary Fig. 2), the average growth rate during the last three hours of the light period was subtracted from the average growth rate during the first three hours in the dark and then averaged across seedlings. Larger values of this 'dark responsiveness index' indicate larger growth rate increases in response to darkness. R scripts are available on request.

RNA extraction and microarray analysis. Our experimental design is based on the observed growth patterns of wild type and *CCA1*-OX under 4L:4D cycles. We reasoned that expression of genes of interest would correlate with growth patterns in the dark. Plants were grown in SD/3 conditions (see above) and whole seedlings were collected 280, 1,240, 1,720, 2,680, 3,160 and 4,120 min after dawn on the fifth day (Supplementary Fig. 5). These time points occur 2 h after lights were turned off in the first and third dark period of each day, over three days. Col grows during the first but not the third dark period, so time points 280, 1,720 and 3,160 were coded as growing (G) for Col and were treated as biological replicates. Time points 1,240, 2,680 and 4,120 were coded as non-growing (NG) for Col and were treated as replicates. All time points were coded as growing for *CCA1*-OX and were treated as replicates. 100 mg frozen tissue samples were ground by pestle and electric motor. Total RNA was extracted using RNeasy Plant Mini kit (Qiagen). From 5 μg of total RNA, complementary RNA was made and labelled with biotin (Affymetrix) according to the protocol used in ref. 33. 15 μg of the resulting fragmented cRNA was hybridized to Affimetrix ATH1 genome array and images were taken by the Affymetrix GeneChip 3000 Scanner. Hybridization and scanning were done at the UC Davis School of Medicine Microarray Core Facility (http://www.ucdmc.ucdavis.edu/medmicro/microarray.html). Robust multi-array averaging[34] and rank product analysis[35] were performed using the affy and RankProd packages, respectively, in Bioconductor /R[36]. Two contrasts were made by Rank Product: first, Col G versus Col NG (growing versus non-growing conditions) and, second, CCA1 G versus Col NG (also growing versus non-growing). Genes were designated as being significantly up- or down-regulated if they had a false-discovery rate value less than 0.1 for both of these contrasts. Gene annotations are based on the latest version of the *Arabidopsis* genome (TAIR6) at TAIR (http://www.arabidopsis.org/) and published literature.

RNA extraction and qRT–PCR analysis. Total RNA was extracted from about 10–20 seedlings using RNeasy Plant Mini kit (Qiagen). The RNA was treated with DNaseI on columns (Qiagen), and 500 ng of eluted RNA was used for complementary DNA synthesis using iScript (Biorad) for samples grown in short-day conditions or SuperScript III reverse transcriptase (Invitrogen) with custom-made oligo-dT (18 nucleotides T). 4 μl of 40-fold diluted cDNA was added to 25 μl PCR buffer containing SYBR Green I and TAQ polymerase (IQ SYBR Green Supermix, BioRad), and fluorescence was detected using an iCycler (Biorad). Primers and PCR conditions for *PP2A* subunit (At1g13320), *PIF4* and *PIF5* were as described in refs 37, 38, except that 50 cycles of two-step PCR were used. Primers for *CCA1*, *TOC* and *GIGANTEA* (*GI*) are described in ref. 39. Data were obtained from two replicate samples in each of two duplicate experiments. The expression of each gene within each sample was normalized against *PROTEIN PHOSPHATASE 2A* (*PP2A*) subunit, and expressed relative to a calibrator sample with the use of the formula $2^{-\Delta\Delta CT}$ as described in ref. 40. ΔΔCT is defined as: [CT gene of interest (unknown sample)–CT *PP2A* (unknown sample)]–[CT gene of interest (calibrator sample)–CT *PP2A* (calibrator sample)], where CT denotes the threshold cycle for detection of PCR product. The expression of *PP2A* does not vary significantly across public microarray data[38]. The calibrator sample was designated as the most highly expressed time point for each gene of interest, and therefore has an expression of 1.0.

Western blot analysis. Generation of *Arabidopsis* plants overexpressing HA-tagged PIF4 or PIF5 protein will be described elsewhere (S.L., P.D.D. and C.F., unpublished). Methods for detecting PIF4–HA or PIF5–HA protein levels are described in ref. 41, except that anti-HA-peroxidase (Roche) was used at 1:1,000 dilution. SuperSignal West Femto Maximum Sensitivity Substrate (Pierce) was used for the peroxidase substrate to detect signals.

30. Strayer, C. *et al.* Cloning of the *Arabidopsis* clock gene *TOC1*, an autoregulatory response regulator homolog. *Science* **289**, 768–771 (2000).
31. Doyle, M. R. *et al.* The *ELF4* gene controls circadian rhythms and flowering time in *Arabidopsis thaliana*. *Nature* **419**, 74–77 (2002).
32. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2005).
33. Schmid, M. *et al.* A gene expression map of *Arabidopsis thaliana* development. *Nature Genet.* **37**, 501–506 (2005).
34. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
35. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573**, 83–92 (2004).
36. Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

37. Czechowski, T., Bari, R. P., Stitt, M., Scheible, W.-R. & Udvardi, M. K. Real-time RT–PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* **38,** 366–379 (2004).
38. Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K. & Scheible, W.-R. Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol.* **139,** 5–17 (2005).
39. Mockler, T. C. *et al.* Regulation of flowering time in *Arabidopsis* by K homology domain proteins. *Proc. Natl Acad. Sci. USA* **101,** 12759–12764 (2004).
40. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25,** 402–408 (2001).
41. Duek, P. D., Elmer, M. V., van Oosten, V. R. & Fankhauser, C. The degradation of HFR1, a putative bHLH class transcription factor involved in light signaling, is regulated by phosphorylation and requires COP1. *Curr. Biol.* **14,** 2296–2301 (2004).

LETTERS

Two distinct modes of guidance signalling during collective migration of border cells

Ambra Bianco[1], Minna Poukkula[1]*, Adam Cliffe[1,2]*, Juliette Mathieu[1], Carlos M. Luque[1]†, Tudor A. Fulga[1]† & Pernille Rørth[1,2]

Although directed migration is a feature of both individual cells and cell groups, guided migration has been studied most extensively for single cells in simple environments[1,2]. Collective guidance of cell groups remains poorly understood, despite its relevance for development and metastasis[3]. Neural crest cells and neuronal precursors migrate as loosely organized streams of individual cells[4,5], whereas cells of the fish lateral line[6,7], *Drosophila* tracheal tubes and border-cell clusters[8] migrate as more coherent groups. Here we use *Drosophila* border cells to examine how collective guidance is performed. We report that border cells migrate in two phases using distinct mechanisms. Genetic analysis combined with live imaging shows that polarized cell behaviour is critical for the initial phase of migration, whereas dynamic collective behaviour dominates later. PDGF- and VEGF-related receptor and epidermal growth factor receptor act in both phases, but use different effector pathways in each. The myoblast city (Mbc, also known as DOCK180) and engulfment and cell motility (ELMO, also known as Ced-12) pathway is required for the early phase, in which guidance depends on subcellular localization of signalling within a leading cell. During the later phase, mitogen-activated protein kinase and phospholipase Cγ are used redundantly, and we find that the cluster makes use of the difference in signal levels between cells to guide migration. Thus, information processing at the multicellular level is used to guide collective behaviour of a cell group.

Border cells perform a well-defined, invasive and directional migration during *Drosophila* oogenesis[8,9]. They delaminate from the follicular epithelium at the anterior end of an egg chamber and migrate posteriorly, towards the oocyte, as a compact cluster (Fig. 1a). They then migrate dorsally towards the oocyte nucleus. The border-cell cluster consists of about six outer migratory border cells and two inner polar cells that induce migratory behaviour in the outer cells but seem to be non-migratory. Two receptor tyrosine kinases (RTKs), PDGF- and VEGF-related receptor (PVR) and epidermal growth factor receptor (EGFR), are guidance receptors for border cells. Both receptors act redundantly during posterior migration towards the oocyte[10], whereas EGFR and its dorsally localized ligand, Gurken, are essential for dorsal migration[11]. Localized signalling from the RTKs is important and actively maintained, especially early in migration[12]. Rac and the atypical Rac exchange factor Mbc (myoblast city, also known as DOCK180) are important effectors[10]. To determine the contribution of Mbc and related proteins (Supplementary Fig. 1), we generated a loss-of-function allele of their common cofactor ELMO (engulfment and cell motility, also known as Ced-12) (ref. 13) by homologous recombination (Fig. 1c). Clusters of *elmo* mutant border cells arrested early in migration, a defect that could be rescued by expressing *elmo* complementary DNA (Fig. 1b, d). As for *mbc*[10], reduction in *elmo* function suppressed F-actin accumulation caused by constitutive PVR signalling, placing ELMO downstream of the receptor in this respect (Supplementary Fig. 1).

To determine whether later steps in migration also depend on ELMO, we investigated mosaic border-cell clusters consisting of wild-type and mutant cells. If a mutation does not affect migration, mutant cells should be distributed randomly within the cluster (Fig. 1e–k). Mutant cells defective in migration would be in the rear, 'carried along' by normal cells[14]. As expected, *Pvr* and *Egfr* double mutant cells were in the rear during posterior migration, as were *Egfr* mutant cells during dorsal migration (Fig. 1k), reflecting the requirements at each stage. *elmo* mutant cells were in the rear during the

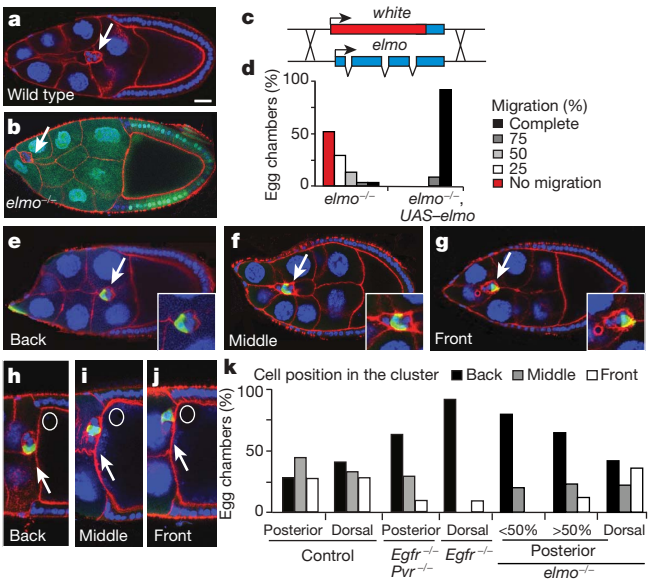


Figure 1 | ELMO is essential in early but not in late border-cell migration. **a**, Shown is a wild-type stage nine egg chamber (anterior is to the left). Blue indicates DNA, red indicates F-actin and the arrow indicates border cells. Scale bar, 20 μm. **b**, *elmo* mutant cluster marked by the absence of GFP (green). **c**, *elmo* locus and knockout strategy. *white* is a gene that inserted in *white*$^{-/-}$ flies changes the eye colour from white to red. **d**, Posterior migration of full *elmo* mutant clusters (n = 31) and rescue by *UAS-elmo* ($n = 12$). **e–j**, Examples of labelled cells (green) within mosaic clusters, in the back, middle or front position during posterior (**e–g**) or dorsal (**h–j**) migration. The circle indicates the oocyte nucleus (dorsal). **k**, Scoring positions of cells with indicated genotypes within mosaic clusters ($15 < n < 32$).

[1]European Molecular Biology Laboratory, Heidelberg, 69117, Germany. [2]Temasek Life Sciences Laboratory (TLL), 1 Research Link, National University of Singapore, Singapore 117604. †Present addresses: Centro Andaluz de Biología del Desarrollo CSIC/UPO, Carretera de Utrera, Km. 1, 41013 Sevilla, Spain (C.M.L.); Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 02115 Boston, USA (T.A.F.).
*These authors contributed equally to this work.

initial migration, but were equally frequent in the leading position during dorsal migration (Fig. 1k). This indicates that, although ELMO is essential for the early-phase signalling, the later phase of migration does not require the Mbc–ELMO complex.

To understand late guidance signalling, we dissected EGFR signalling, on which dorsal migration depends. Uniformly activated EGFR, like PVR, dominantly impairs migration[11]. The carboxy-terminal tail of EGFR was essential for this activity. Systematic mutagenesis of all docking tyrosines to phenylalanine identified Y1357 as being critical, with minor contributions from Y1405 and Y1406 (Fig. 2a, b). Other tyrosines, including Y1095 in the conserved activation loop (phosphorylated in HER2 (Human EGF Receptor 2), ref. 15), were not required. Twenty Src-homology 2- and phosphotyrosine-binding-containing signalling molecules were tested for binding to active EGFR and tyrosine mutants (Supplementary Table 1). Y1357 was necessary and sufficient for binding of the adaptor protein Shc and its phosphotyrosine-binding domain. No other tested interactor behaved in this way. Binding was confirmed by immunoprecipitation (Supplementary Fig. 2). Border cells mutant for *Shc* showed no dorsal migration (Fig. 2c) and, when PVR signalling was also blocked, these cells showed severely impaired posterior migration (Fig. 2d, e). This phenotype is identical to that of *Egfr* mutant cells, suggesting that Shc is essential immediately downstream of EGFR for guidance signalling.

The Shc adaptor protein links EGFR and other RTKs to mitogen-activated protein kinase (MAPK) kinase signalling as well as to other classical downstream pathways[16]. Raf, phospholipase Cγ (PLC-γ) or phosphatidylinositol-3-OH-kinase are not uniquely required for migration[11]; however, the pathways might act redundantly. Simultaneous perturbation of PLC-γ and Raf impaired migration, with no effect of phosphatidylinositol-3-OH kinase (Supplementary Fig. 3). Double mutant border-cell clusters, cell-autonomously lacking PLC-γ and Raf or lacking PLC-γ and MAPK kinase (MAPKK), initiated migration but were delayed later in posterior migration (Fig. 2f, g) and showed no dorsal migration. This phenotype is more severe than that of *Egfr* or *Shc* alone, suggesting that both RTKs might be affected. Prevention of PVR activity in double mutant cells did not block posterior migration (Fig. 2f), confirming that the requirement for these pathways was stage-specific and not EGFR-specific. Finally,

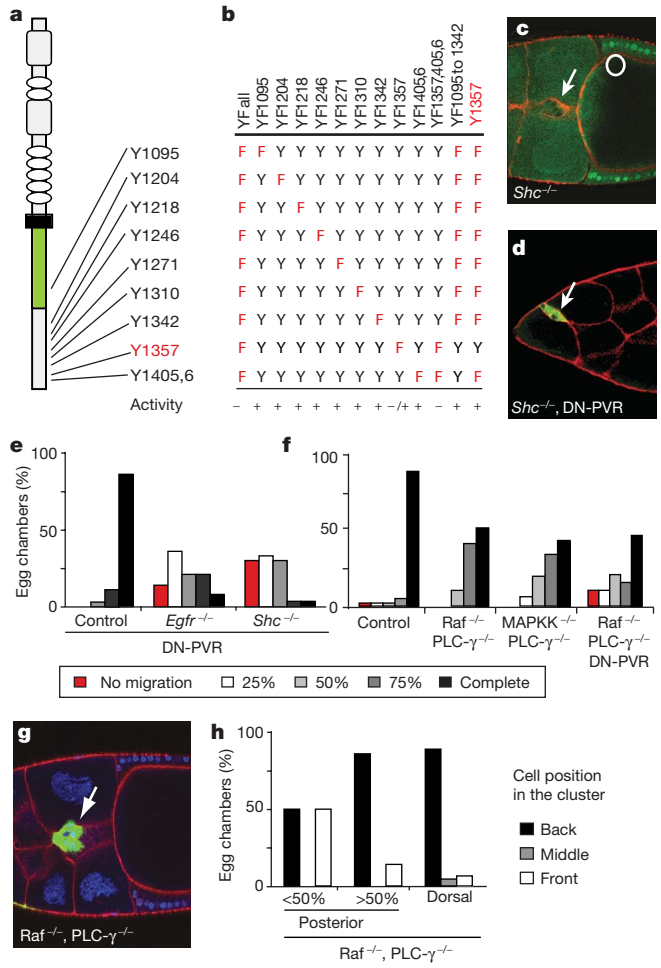


Figure 2 | Shc is downstream of EGFR, and PLCγ and Raf/MAPK are required for late RTK signalling. **a**, Shown is EGFR, SwissProt entry P04412. **b**, Y-to-F mutants. +, active (as *UAS-λ-Egfr* with *slbo-Gal4*); –, no effect; and –/+, intermediate. **c**, Stage-ten $Shc^{111\text{-}40}$ mutant cluster (GFP negative). All migrated posteriorly; none (0 of 6) migrated dorsally. The circle represents the oocyte nucleus. **d**, $Shc^{111\text{-}40}$ mutant clone expressing dominant negative (DN)-PVR (GFP positive). **e**, **f**, Quantification of posterior migration by clusters of indicated genotypes; $42 < n < 52$. MAPKK = $Dsor^{LH110}$, Raf = phl^{11}. None but the control migrated dorsally. **g**, PLC-γ (sl^{1}) Raf double mutant cluster (GFP positive). **h**, Distribution of double mutant cells in mosaic clusters; $n = 8$–45. Cell position in cluster: black bars, back; grey bars, middle; white bars, front.

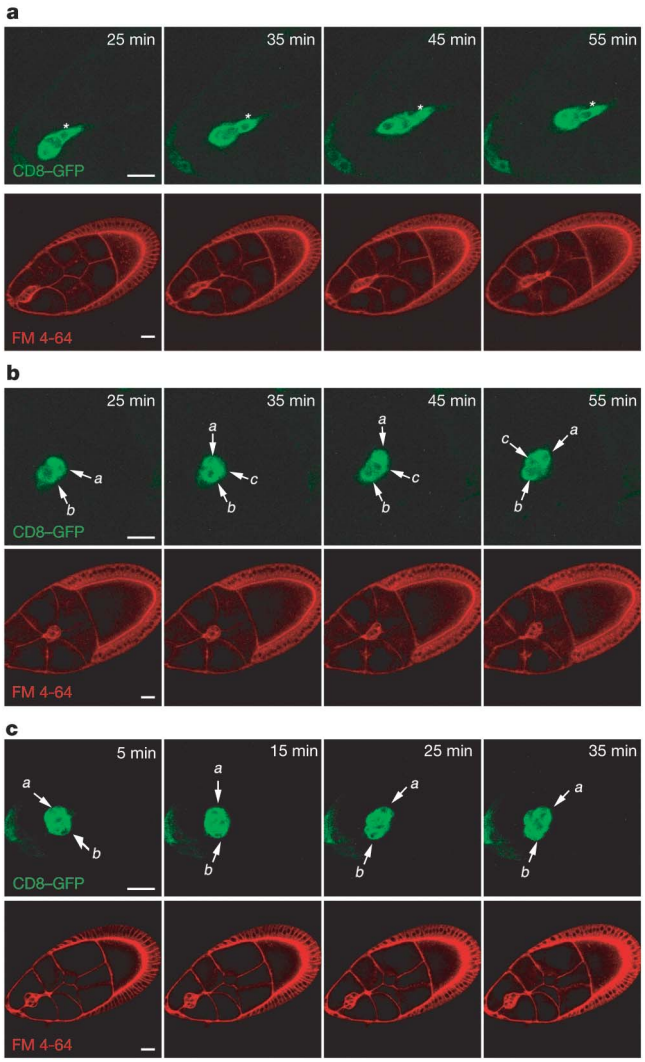


Figure 3 | Live imaging of border-cell migration. **a**, Early migration (stills of Supplementary Movie 1). Cluster migrates rapidly (~1 μm min^{-1}) with elongated, polarized morphology. The asterisk marks the leading cell. CD8–GFP marks border cells; membranes are labelled with FM 4-64. **b**, Late migration, demonstrating shuffling (stills of Supplementary Movie 4). Individual cells (marked *a*–*c*) change position within the cluster (arrows). **c**, PVF1 overexpression (stills of Supplementary Movie 6). Morphology and movement (slow forward movement) is similar to wild-type late phase. GFP images are a projection of four Z-sections; FM 4-64 images are a single confocal section corresponding to the centre of the cluster. Scale bar, 20 μm.

analysis of mosaic clusters showed that Raf/MAPK and PLC-γ were important in late migration (Fig. 2h), reciprocal to the requirement for *elmo* (Fig. 1k). These results genetically define two migratory phases: an early posterior phase requiring ELMO–Mbc and a later posterior and dorsally directed phase requiring Raf/MAPK or PLC-γ. Both RTKs shift effector-pathway-dependency as migration progresses.

To investigate the different migratory phases, we examined border-cell migration via live imaging. We first established appropriate conditions for culturing and imaging of egg chambers (see Methods), considering only active, growing ones. Border cells were selectively labelled with green fluorescent protein (GFP) and all membranes were labelled with the vital dye FM4-64 (Fig. 3 and Supplementary Movies). For all 24 wild-type samples, the identity of the front cell changed during the observation period, confirming the inference from fixed samples that cells change position during migration (Fig. 1k). This indicates that there is no determined front-cell fate. We observed a clear difference in behaviour of clusters during early (first half) and late phases. Early clusters had one, sometimes two, highly polarized cells clearly leading the migration; once these cells delaminated they moved straight and relatively fast (1 μm min^{-1}; Fig. 3a and Supplementary Movies 1 and 2). Weakly stained extensions protrude far from delaminating cells (Supplementary Movie 2) and subsequently shorten during movement, suggesting a 'grapple and pull' mechanism[17]. Midway towards the oocyte, strong polarization was lost and cells rounded and started to 'shuffle' while dynamically probing the environment with short extensions. Occasionally the cluster would rotate or 'tumble' completely (Fig. 3b and Supplementary Movies 3 and 4). This shuffling behaviour still provided effective movement of the cluster towards the oocyte and dorsally, albeit more slowly (0.4 μm min^{-1}). Labelling cells with nuclear GFP (Supplementary Movie 7) allowed visualization of changes in positions within the cluster. The front cell exchanged, on average, every 18 min.

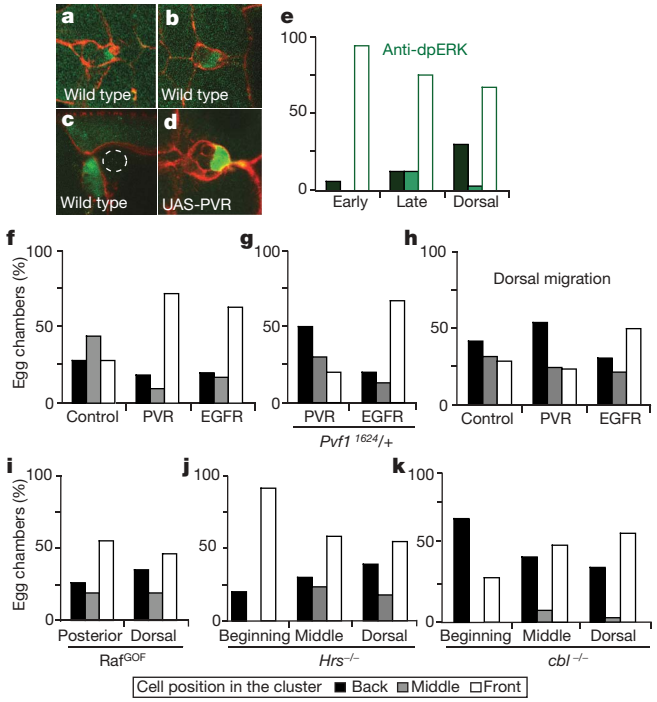


Figure 4 | Guidance of border-cell clusters by collective decision comparing signal levels between cells. **a–c**, Anti-DpErk staining in wild-type egg chambers and **d**, with front two cells overexpressing PVR (mosaic cluster). **e**, Position of high anti-DpErk staining cell(s) within wild-type clusters; $16 < n < 33$. Black bars, strong staining even throughout cluster; green bars, strong staining at back; white bars, strong staining at front. **f–h**, Position of PVR- and EGFR-overexpressing cells in mosaic clusters during posterior migration (**f**, $32 < n < 53$), during posterior migration in $Pvf1^{1624}/+$ background (**g**, $10 < n < 30$) and during dorsal migration (**h**, $23 < n < 49$). **i**, Position of cells expressing Raf-GOF (posterior and dorsal migration, $64 < n < 74$). **j**, **k**, Position of *Hrs* (**j**, $13 < n < 74$) and *Cbl* (**k**, $10 < n < 39$) mutant cells in mosaic clusters.

As expected, positions corresponding to the second, slower phase of migration were more represented when cluster position along the migratory path was quantified in fixed samples (Supplementary Fig. 4). Also, border cells expressing dominant negative PVR and EGFR were individually active but provided little net cluster movement (Supplementary Movie 5), as expected from the lack of guidance information[10]. Finally, uniform overexpression of the attractant PVF1 caused an increased shuffling behaviour in the early phase (Fig. 3c and Supplementary Movies 6 and 8) but allowed slow forward movement, resembling normal late migration. This indicates that migrating clusters can interpret a shallow gradient when using the shuffling mode (Supplementary Movie 8). It also suggests that the normal change in migratory behaviour midway into posterior migration might be triggered by the higher concentration of ligands closer to the oocyte.

The early phase of migration with a highly polarized front cell corresponds temporally to the genetic requirement for ELMO activity (Fig. 1). During the later phase, individual *elmo* mutant cells can alternate with wild-type cells in the lead position (Supplementary Movie 11). Genetic analysis showed that Raf and MAPKK and, by inference, MAPK activation was sufficient to convey late guidance information. This was puzzling because MAPK activation appeared uniform in migrating border cells[11] (Fig. 4a–d), and localized effects are usually a hallmark of guidance signalling. However, signalling that is not localized within an individual cell could still transmit spatial guidance information to the cell cluster if the cell with higher overall signalling indicates the direction of subsequent migration for the whole cluster, as observed for MAPK signalling in border cells (Fig. 4a–c, quantified in Fig. 4e). In this 'collective guidance' scenario, each cell of the cluster can be thought of as being analogous to a sector of an individual guided cell. Different levels of signalling in individual cells of the cluster transform into migration vectors because border cells adhere to each other and these contacts differ from substrate contacts (model in Supplementary Fig. 5). The occasional tumbling of border-cell clusters emphasizes the ability of these cells to behave as a collective unit. Tumbling may help single guided cells to 'reassess' their environment[18].

To test this model for guidance, we manipulated the relative levels of signalling in individual cells of the cluster. Dynamic shuffling should allow cells to constantly 'compete' for the front position. None of the manipulations discussed below improved migration if all cells in a cluster were affected. Individual border cells with moderately elevated levels of PVR or EGFR were preferentially in the front relative to wild-type cells (Fig. 4d, f). Cells with elevated PVR tended to stay in or near the front position (Supplementary Movie 9), suggesting that they were not competed away by other cells. This bias was ligand-dependent, because reducing PVF1 levels shifted the bias from PVR to EGFR, as was also shown by analysis of dorsal migration (Fig. 4g, h). Thus, increased signalling gives a cell-front bias when measuring an informative ligand. Elevating intracellular signalling levels had similar effects, whether by overexpression of an active form of Raf (Fig. 4i) or by preventing downregulation of signalling as in *Hrs* mutant cells (Fig. 4j), in which RTK-mediated MAPK signalling is elevated in enlarged endosomes[19]. The more modest front bias in *Hrs* mutant cells was reflected in behaviour: they could be displaced from the front (Supplementary Movie 10). The E3 ubiquitin ligase Cbl negatively regulates RTK signalling[20] and is also required to maintain localized RTK signalling within border cells initiating migration[12]. *Cbl* mutant cells shifted from being preferentially at the back during early stages to being in the front during later migration (Fig. 4k). This indicates a transition from a mode requiring Cbl-dependent localization of signalling within the leading cell to a mode

based on collective decisions within the cluster, in which *Cbl* mutant cells have an advantage owing to elevated RTK signalling.

We suggest that guidance of border-cell migration is achieved by two means: signalling localized within the cell, as used in individual migrating cells[1,2], and collective guidance, whereby the cluster uses differences in signalling strength among its constituent cells to determine direction. The two modes use the same guidance cues and receptors, but different downstream effectors. Localized signalling is required for the initial, polarized rapid migration, whereas collective behaviour, though observable throughout, dominates in the later phase. Collective decisions on the basis of differences in RTK signalling strength are important in *Caenorhabditis elegans* vulval development[21] and in branching of *Drosophila* tracheal tubes[22], in which they result in specification of discrete cell fates. This differs from the dynamic situation reported here, in which the identity of the leading cell constantly changes. Indeed, the frequent exchange of leading cells suggests that front behaviour is normally temporarily restricted, possibly by induced inactivation of signalling. Such dynamics may allow the cluster to better reassess the environment. For guided migration of cell groups, our analysis indicates that sensing and regulation happens both at the single cell level and at the next level—that of collective cell decisions.

METHODS SUMMARY

The $elmo^{KO}$ mutant (a null allele) was generated by homologous recombination[23]. All migration analysis and scoring was done as clonal analysis, scoring egg chambers in which all (complete clone) or only one to three (mosaic cluster) border cells were mutant. Genotypes used for Shc: *hsFLP/+;; $Shc^{111\text{-}40}$ FRT80/ubi-GFP, FRT80;* and *hsFLP, UAS-LacZ; slbo-Gal4, UAS-DN-PVR/+; $Shc^{111\text{-}40}$ FRT8/tub-Gal80, FRT80*. Genotypes used for Raf and MAPKK: phl^{11}, sl^{1}, *FRT19/tub-Gal80, FRT19; UAS-GFP, hsFLP/+; UAS-p35/+* and $Dsor^{LH110}$, sl^{1}, *FRT19/tub-Gal80, FRT19; UAS-GFP, hsFLP/+; UAS-p35/+*. *UAS-p35* prevents apoptosis and allows recovery of more mutant clones; the control in Fig. 2f was therefore: *FRT19/tub-Gal80, FRT19; UAS-GFP, hsFLP/+; UAS-p35/+*. For Fig. 4j, k, the genotypes *hsFLP/+; Hrs^{D28}, FRT40/ubi-GFP, FRT40* and *hsFLP/+;; Cbl^{F165},FRT80/ubi-GFP, FRT80* were used. For Fig. 4e–i, the genotypes used were: *UAS-lacZ, hs-FLP/+; FRT40,tub-Gal80/FRT40; c522-Gal4/UAS-X*, where X is PVR, EGFR or Raf-GOF (gain of function) (ref. 24, F179), with LacZ marking overexpressing cells. For Fig. 4b the genotype was: *UAS-lacZ, hs-FLP/ $Pvf1^{1624}$; FRT40,tub-Gal80/FRT40; c522-Gal4/UAS-X*. c522-Gal4 is a weaker border-cell driver than *slbo-Gal4* (ref. 25) and c522-Gal4-expressing cells could not be reliably identified at very early stage nine.

Point mutants (Y to F) in the carboxy-terminal region of EGFR were generated, cloned into pGBK and pRm-Flag, and tested for binding to preys tagged with haemagglutinin at the amino terminus by two-hybrid analysis and immunoprecipitation. *UAS-λ*-Egfr mutant-carrying transgenics were crossed to *slbo-Gal4, $slbo^{1310\text{-}lacZ}$/+* flies and were scored for migration delay at stage ten.

For live imaging, yeast-fed females were dissected in Schneider's media plus 5 μg ml^{-1} insulin, and imaged in dissection media supplemented with 2.5% fetal calf serum, 2 mg ml^{-1} trehalose, 5 μM methoprene, 1 μg ml^{-1} 20-hydroxyecdysone, 50 ng ml^{-1} adenosine deaminase and 9 μM FM 4-64. Multi-position imaging was performed with a Zeiss Meta confocal microscope. Dissection time did not exceed 15 min, and imaging was carried out for no longer than 2.5 h after the start of dissection.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 April; accepted 25 May 2007.

1. Parent, C. A. & Devreotes, P. N. A cell's sense of direction. *Science* **284,** 765–770 (1999).
2. Servant, G. *et al.* Polarization of chemoattractant receptor signaling during neutrophil chemotaxis. *Science* **287,** 1037–1040 (2000).
3. Friedl, P., Hegerfeldt, Y. & Tusch, M. Collective cell migration in morphogenesis and cancer. *Int. J. Dev. Biol.* **48,** 441–449 (2004).
4. Kulesa, P. M. & Fraser, S. E. Neural crest cell dynamics revealed by time-lapse video microscopy of whole embryo chick explant cultures. *Dev. Biol.* **204,** 327–344 (1998).
5. Lois, C., Garcia-Verdugo, J. M. & Alvarez-Buylla, A. Chain migration of neuronal precursors. *Science* **271,** 978–981 (1996).
6. Ghysen, A. & Dambly-Chaudiere, C. Development of the zebrafish lateral line. *Curr. Opin. Neurobiol.* **14,** 67–73 (2004).
7. Haas, P. & Gilmour, D. Chemokine signaling mediates self-organizing tissue migration in the zebrafish lateral line. *Dev. Cell* **10,** 673–680 (2006).
8. Starz-Gaiano, M. & Montell, D. J. Genes that drive invasion and migration in *Drosophila. Curr. Opin. Genet. Dev.* **14,** 86–91 (2004).
9. Rørth, P. Initiating and guiding migration: lessons from border cells. *Trends Cell Biol.* **12,** 325–331 (2002).
10. Duchek, P., Somogyi, K., Jékely, G., Beccari, S. & Rørth, P. Guidance of cell migration by the *Drosophila* PDGF/VEGF receptor. *Cell* **107,** 17–26 (2001).
11. Duchek, P. & Rørth, P. Guidance of cell migration by EGF receptor signaling during *Drosophila* oogenesis. *Science* **291,** 131–133 (2001).
12. Jekely, G., Sung, H. H., Luque, C. M. & Rorth, P. Regulators of endocytosis maintain localized receptor tyrosine kinase signaling in guided migration. *Dev. Cell* **9,** 197–207 (2005).
13. Brugnera, E. *et al.* Unconventional Rac-GEF activity is mediated through the Dock180–ELMO complex. *Nature Cell Biol.* **4,** 574–582 (2002).
14. Rørth, P., Szabo, K. & Texido, G. The level of C/EBP protein is critical for cell migration during *Drosophila* oogenesis and is tightly controlled by regulated degradation. *Mol. Cell* **6,** 23–30 (2000).
15. Bose, R. *et al.* Phosphoproteomic analysis of Her2/neu signaling and inhibition. *Proc. Natl Acad. Sci. USA* **103,** 9773–9778 (2006).
16. Luschnig, S., Krauss, J., Bohmann, K., Desjeux, I. & Nüsslein-Volhard, C. The *Drosophila* SHC adaptor protein is required for signaling by a subset or receptor tyrosine kinases. *Mol. Cell* **5,** 231–241 (2000).
17. Fulga, T. A. & Rørth, P. Invasive cell migration is initiated by guided growth of long cellular extensions. *Nature Cell Biol.* **4,** 715–719 (2002).
18. Reichman-Fried, M., Minina, S. & Raz, E. Autonomous modes of behavior in primordial germ cell migration. *Dev. Cell* **6,** 589–596 (2004).
19. Lloyd, T. E. *et al.* Hrs regulates endosome membrane invagination and tyrosine kinase receptor signaling in *Drosophila. Cell* **108,** 261–269 (2002).
20. Pai, L. M., Barcelo, G. & Schupbach, T. *D-cbl*, a negative regulator of the Egfr pathway, is required for dorsoventral patterning in *Drosophila* oogenesis. *Cell* **103,** 51–61 (2000).
21. Yoo, A. S., Bais, C. & Greenwald, I. Crosstalk between the EGFR and LIN-12/Notch pathways in *C.elegans* vulval development. *Science* **303,** 663–666 (2004).
22. Ghabrial, A. S. & Krasnow, M. A. Social interactions among epithelial cells during tracheal branching morphogenesis. *Nature* **441,** 746–749 (2006).
23. Gong, W. J. & Golic, K. G. Ends-out, or replacement, gene targeting in *Drosophila. Proc. Natl Acad. Sci. USA* **100,** 2556–2561 (2003).
24. Ghiglione, C. *et al.* The transmembrane molecule Kekkon 1 acts in a feedback loop to negatively regulate the activity of the *Drosophila* EGF receptor during oogenesis. *Cell* **96,** 847–856 (1999).
25. Borghese, L. *et al.* Systematic analysis of the transcriptional switch inducing migration of border cells. *Dev. Cell* **10,** 497–508 (2006).

Supplementary information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank EMBL-ALMF, K. Somogyi and A. M. Voie for help, Y. Cohen and E. Schejter for sharing, and K. Brown, V. Hietakangas, D. Gilmour and S. Cohen for comments. This work was supported by Marie Curie FP6 Intra-European fellowships (A.C. and M.P.), the Academy of Finland and the Helsingin Sanomat Centennial Foundation (M.P.), HFSP (J.M.), EMBO (C.M.L.) and EMBL.

Author Contributions M.P. and A.C. contributed equally to this work and performed all culturing and live imaging experiments. J.M. contributed to the *elmo* knockout. C.M.L. and T.A.F. performed the protein interaction and genetic analysis of EGFR, respectively. A.B. performed all remaining analyses. P.R. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.R. (Pernille@tll.org.sg).

METHODS

ELMO. The *elmo* knockout allele $elmo^{KO}$ was generated by homologous recombination[23], resulting in a 2,062 bp deletion (2L:12097872 to 12099934, Flybase). $elmo^{KO}$ was tested by PCR and by genetic analysis: $elmo^{KO}$ is homozygous lethal, fails to complement deficiencies for the region, and the defect in border cells is rescued by a wild-type *elmo* cDNA[26] cloned into pUAST. Complete mutant clones were recovered four days after adult heat shock of the genotype: *hsFLP; $elmo^{KO}$, FRT40/ubi-GFP,FRT40* (mutant) and *hsFLP,UAS-CD8GFP; tubGal80,FRT40/ $elmo^{KO}$,FRT40; tub-Gal4/UAS-Elmo* (rescue). For comparison of *elmo* mutant phenotypes with those of *mbc*-related genes, see Supplementary Fig. 1.

EGFR. Individual point mutations (Y to F) were introduced into the C-terminal region of *Egfr* by PCR and were combined by standard subcloning or multiple rounds of mutagenesis. Mutants were subcloned into *UAS-λ-Egfr* for analysis, into pGBK-EGFRi (complete intracellular domain of EGFR, auto-phosphorylation-competent in yeast) for yeast two-hybrid analysis, and into pRm-EGFR-Flag for transfection. Preys were cloned into pGADT7 by linker-PCR with *Bam*HI-*Xho*I, and some into pRm with haemagglutinin tagging at the N terminus. Yeast two-hybrid, Schneider cell transfection, immunoprecipitation and western blotting were performed using standard procedures. Phenotypes of EGFR mutants in the context of *UAS-λ*-Egfr (*UAS-λ-top*) were scored in stage-ten egg chambers as described in ref. 11, with *slbo-Gal4, $slbo^{1310\text{-}lacZ}$/+*. At least three independent lines were tested per mutant construct. UAS-λ-EGFR-ΔC (including the kinase domain but truncated at amino acid 1202) had no effect in this assay.

Additional clonal analysis. Complete mutant clones were recovered four to five days after adult heat shock of the following genotypes: for Shc, *hsFLP/+;;$Shc^{111\text{-}40}$ FRT80/ubi-GFP, FRT80*; and *hsFLP, UAS-LacZ; slbo-Gal4, UAS-DN-PVR/+; $Shc^{111\text{-}40}$ FRT80/ tub-Gal80, FRT80*. For Raf and MAPKK, *phl^{11}, sl^{1}, FRT19/tub-Gal80, FRT19; UAS-GFP, hsFLP/+; UAS-p35/+* and *$Dsor^{LH110}$, sl^{1}, FRT19/tub-Gal80, FRT19; UAS-GFP, hsFLP/+; UAS-p35/+*. UAS-p35 is used to prevent apoptosis and to allow recovery of more mutant clones; the control in Fig. 2f was therefore *FRT19/tub-Gal80, FRT19; UAS-GFP, hsFLP/+; UAS-p35/+*. Mixed mutant clones (one to three mutant border cells) were recovered by adult heat shock of the same genotypes two to three days before dissection. For Fig. 4j, k, the genotypes *hsFLP/+; Hrs^{D28}, FRT40/ubi-GFP, FRT40* and *hsFLP/+;; Cbl^{F165},FRT80/ubi-GFP, FRT80* were used. For small clones where only one or two cells are overexpressing the indicated constructs (Fig. 4f, h, i), females of the genotype *UAS-lacZ, hs-FLP/+; FRT40,tub-Gal80/FRT40; c522-Gal4/UAS-X*, where X is PVR, EGFR or Raf-GOF (ref. 24, F179), were heat shocked and expressing cells were identified by anti-β-galactosidase staining. For Fig. 4g, the genotype was *UAS-lacZ, hs-FLP/$Pvf1^{1624}$; FRT40,tub-Gal80/FRT40; c522-Gal4/UAS-X*; c522-Gal4 is a weaker border-cell-specific driver than slbo-Gal4 (ref. 25). c522-Gal4-expressing cells could not be determined reliably at very early stage nine.

Screening of culture conditions. To identify suitable culture conditions to image border-cell migration, we developed a high-throughput imaging screen. In summary, egg chambers were dissected for 1 h in media, and then appropriate stages were sorted using a fluorescence dissecting microscope. Groups of egg chambers were transferred to a 96-well plate (Nunc, 167008), with each well containing 100 μl test media. We used an Olympus Cell multi-position fluorescence microscope to perform time-lapse imaging. Typically, GFP and transmission images of 60–100 egg chambers (in 6–10 media conditions) were acquired with a ×20 objective at 30 min time intervals for 12 h. Different media and supplement combinations were assessed according to their effects on border-cell migration as well as their overall morphology and development of the egg chambers. More detailed information on media conditions tested and other aspects of live imaging of border cells is available on request.

Preparation of egg chambers for confocal imaging. Egg chambers were imaged in eight-well Lab-Tek chambers (Nunc, 155411). Chambers were freshly coated for 2 h with 0.1 mg ml^{-1} poly-D-lysine (>300 kDa in PBS, Sigma, P-7405), and then washed four times with water before imaging.

Females were taken three to six days after hatching and were fed fresh yeast one day before dissection. Dissections were performed in Schneider's medium (Gibco 21720) plus 5 μg ml^{-1} insulin (Sigma, 19278). Ovaries were removed from one to two females, washed in dissection medium, dissociated, and egg chambers of appropriate stages were removed from the muscle sheath using fine dissection pins. Appropriate egg chambers were washed in medium, moved to imaging chambers and overlaid with 200 μl imaging medium: Schneider's medium plus 2.5% fetal calf serum (PAA, A15043), 5 μg ml^{-1} insulin, 2 mg ml^{-1} trehalose (Fluka, 90208), 5 μM methoprene (Sigma, 33375), 1 μg ml^{-1} 20-hydroxyecdysone (Sigma, H5142), 50 ng ml^{-1} adenosine deaminase (Roche, 10258921) and 9 μM FM 4-64 (Invitrogen, T13320).

Total dissection time did not exceed 15 min, and imaging was generally started 30 min after dissection began.

Confocal microscopy imaging. Imaging was performed using a Zeiss LSM-510-Meta with a ×40 1.3 N.A. Plan NeoFluor oil immersion objective. Three channels were acquired simultaneously: GFP (488 nm laser and 505–550 band-pass filter), FM 4-64 (488 nm laser and 560 nm long pass filter) and transmission image (DIC). Seven sections were taken 7.5 μm apart with 2–5 min between stacks. Three to six egg chambers were simultaneously imaged using a multi-time series macro[27]. The three to four sections covering the migrating cluster were projected for each time point, using the Zeiss LSM image examiner software.

Egg chambers were chosen in which border cells had clearly delaminated or were already migrating. FM 4-64 was used as a membrane marker and as an indicator of damaged egg chambers. Egg chambers were excluded either before or after imaging if nurse or follicle cells showed greatly increased FM 4-64 uptake compared to their neighbours (indicating damage during dissection). The DIC image was also used to assess the overall health of the egg chamber during imaging. We observed that the nurse-cell nuclei moved and rotated during imaging. After several hours (usually ~3 h post-dissection) nuclei stopped moving, indicating the early stages of degeneration. After 2–3 h, nurse-cell membranes begin detaching from each other, starting at the corners. Movies were not continued after nurse-cell detachment or cessation of nuclear movement was observed. Using these criteria, we found that we were able to reliably image egg chambers for 2.5–3 h after dissection. Movies were brightness- and contrast-adjusted with ImageJ, and the green channels of Supplementary Movies 7–9 were treated with a 1-pixel-radius gaussian filter to reduce background noise.

Determination of migration speed. As border-cell clusters move relatively little in the Z-plane, we tracked their approximate speeds of migration in the *X*/*Y* axis only using the manual tracking plugin for ImageJ (http://rsb.info.nih.gov).

26. Zhou, Z., Caron, E., Hartwieg, E., Hall, A. & Horvitz, H. R. The *C.elegans* PH domain protein CED-12 regulates cytoskeletal reorganization via a Rho/Rac GTPase signaling pathway. *Dev. Cell* **1**, 477–489 (2001).
27. Rabut, G. & Ellenberg, J. Automatic real-time three-dimensional cell tracking by fluorescence microscopy. *J. Microsc.* **216**, 131–137 (2004).

Vol 448|19 July 2007|doi:10.1038/nature05929

LETTERS

The Rab8 GTPase regulates apical protein localization in intestinal cells

Takashi Sato[1], Sotaro Mushiake[2], Yukio Kato[3], Ken Sato[1], Miyuki Sato[1], Naoki Takeda[4], Keiichi Ozono[2], Kazunori Miki[5], Yoshiyuki Kubo[3], Akira Tsuji[3], Reiko Harada[1] & Akihiro Harada[1]

A number of proteins are known to be involved in apical/basolateral transport of proteins in polarized epithelial cells[1–7]. The small GTP-binding protein Rab8 was thought to regulate basolateral transport in polarized kidney epithelial cells through the AP1B-complex-mediated pathway[8,9]. However, the role of Rab8 (Rab8A) in cell polarity *in vivo* remains unknown. Here we show that Rab8 is responsible for the localization of apical proteins in intestinal epithelial cells. We found that apical peptidases and transporters localized to lysosomes in the small intestine of Rab8-deficient mice. Their mislocalization and degradation in lysosomes led to a marked reduction in the absorption rate of nutrients in the small intestine, and ultimately to death. Ultrastructurally, a shortening of apical microvilli, an increased number of enlarged lysosomes, and microvillus inclusions in the enterocytes were also observed. One microvillus inclusion disease patient who shows an identical phenotype to Rab8-deficient mice expresses a reduced amount of RAB8 (RAB8A; NM_005370). Our results demonstrate that Rab8 is necessary for the proper localization of apical proteins and the absorption and digestion of various nutrients in the small intestine.

To determine the function of *Rab8* in cell polarity *in vivo*, we generated *Rab8* conditional knockout mice. We used both the Cre–*loxP* and *FRT*–Flp systems[10], and combined these with a strong splice-acceptor and internal ribosomal entry site sequence[11] to induce a reverting of the mutant phenotype. We named this system 'revertible knockout'; a detailed description is shown in Supplementary Figs 2 and 3. To confirm the efficacy of this system, we performed a western blot analysis of crude extracts from $Rab8^{+/+}$, $Rab8^{geo/geo}$, $Rab8^{flox/flox}$ and $Rab8^{-/-}$ mouse small intestines. As expected, Rab8 was undetectable in $Rab8^{geo/geo}$ mice, but the amount of Rab8 in 'reverted' $Rab8^{flox/flox}$ mice was almost the same as that in $Rab8^{+/+}$ mice (Supplementary Fig. 4). Additionally, although both $Rab8^{-/-}$ and $Rab8^{geo/geo}$ mice died 3–4 weeks after birth, 'reverted' $Rab8^{flox/flox}$ mice were as healthy as $Rab8^{+/+}$ mice (Supplementary Fig. 4). Morphological phenotypes returned to normal in the 'reverted' ($Rab8^{flox/flox}$ or $Rab8^{flox/-}$) mice as well (Supplementary Fig. 5). Therefore, the 'revertible' knockout system worked as expected.

$Rab8^{-/-}$ and $Rab8^{geo/geo}$ mice showed identical phenotypes (thus, we collectively called $Rab8^{-/-}$ and $Rab8^{geo/geo}$ as 'knockouts' or 'mutants'), presented with diarrhoea and progressive wasting from 2.5 to 3 weeks and died within 5 weeks of birth. Their small intestines were swollen and contained undigested milk. Unexpectedly, basolateral markers (LDL receptor (LDL-R/Ldlr) and Na^+,K^+ ATPase (Atp1a1)[12]) showed proper basolateral localization in the intestinal epithelial cells of the mutants (Fig. 1a). On the other hand, apical markers such as dipeptidyl peptidase IV (DPPIV/Dpp4), alkaline phosphatase (AlP), sucrase-isomaltase (SI)[13] and the oligopeptide transporter 1 (PepT1/Slc15a1)[14] markedly decreased in amount in the apical membrane and accumulated intracellularly (Fig. 1b). Interestingly, such abnormal intracellular accumulations started from postnatal 2 to 2.5 weeks after birth (Supplementary Fig. 6a). The intracellular accumulation was strongly positive for the late endosome/lysosome markers, lysosomal-associated membrane protein 2 (Lamp2) and cathepsin D (Ctsd) and negative for the Golgi/TGN markers GS28 (Gosr1)(Fig. 1c and Supplementary Fig. 6b) and TGN38 (Tgoln1). The intracellular accumulation was not positive for autophagosome marker LC3 (Map1lc3a) (ref. 15). These results suggest that the intracellular structures are late endosomes/lysosomes.

In the kidneys of the knockout mice, we observed a slight increase in the number of lysosomes underneath the apical plasma membrane in proximal tubules. However, the localization of basolateral markers was normal. In addition, the distal tubules of knockout mice showed normal distributions of both apical and basolateral markers. We found no difference in the shape of bile canaliculi and the localization of apical markers in the knockout hepatocytes (Supplementary Fig. 7). Thus, Rab8 does not seem to be important in transcytosis in the liver. In the intestinal epithelial cells of wild-type mice, Rab8 localized to sub-apical (Fig. 1a, arrowheads) and perinuclear punctate (Fig. 1a, arrows) structures. Perinuclear Rab8 partially co-localized with the Golgi apparatus and early endosomes[16] (Supplementary Fig. 8). Sub-apical Rab8 co-localized with or was adjacent to early endosomes. Rab8 did not co-localize with Lamp2. We also observed a GFP (green fluorescent protein) fusion protein with a worm Rab8 homologue in sub-apical structures in *Caenorhabditis elegans* (Supplementary Fig. 9). Additionally, in the *rab-8*-deficient worm, an apical marker, GFP–PGP-1, accumulated beneath the apical plasma membrane of the intestine at the L3–L4 stage (Supplementary Fig. 10). This implicates the evolutionarily conserved role of Rab8 as the localization of apical proteins in the intestine.

To determine whether the apical markers are degraded in these organelles, we measured the levels of these markers by western blot analysis. The amounts of DPPIV, sodium/glucose cotransporter (SGLT/Slc5a1), and PepT1 were reduced in crude extracts from the intestines of the mutants at 3 weeks to 32, 37 and 31% of those of the control, respectively (Fig. 2a, b). The amounts of LDL-R and Sec8 (Exoc4) in the knockout intestine are similar to those in the intestine of the wild-type mice (Fig. 2a, b).

To investigate the functional expressions of glucose and oligopeptide transporters in apical membranes of the small intestine, the uptake rates of typical substrates for the transport systems, α-methyl d-glucopyranoside (α-MDG; a substrate for SGLT) and glycylsarcosine

[1]Laboratory of Molecular Traffic, Department of Molecular and Cellular Biology, Institute for Molecular and Cellular Regulation, Gunma University, Gunma 371-8512, Japan. [2]Department of Pediatrics, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan. [3]Division of Pharmaceutical Sciences, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa 920-1192, Japan. [4]Division of Transgenic Technology, Center for Animal Resources and Development (CARD), IRDA, Kumamoto University, Kumamoto 860-0811, Japan. [5]Department of Pediatrics, Itami Municipal Hospital, Hyogo 664-8540, Japan.

(Gly-Sar; a substrate for PepT1), respectively, were examined[17]. At two weeks of age, the uptake rates of [^{14}C]α-MDG were almost comparable between the two genotypes; they decreased in the presence of unlabelled α-MDG (Fig. 2e, left panel, and Supplementary Fig. 11). Thus, the saturable uptake mechanism for α-MDG from apical membranes was not affected by knocking out *Rab8* at this age. At three weeks of age, on the other hand, the uptake rate of [^{14}C]α-MDG in the small intestines of the mutant mice was much lower than that observed in the control mice. In addition, the uptake rates of [^{14}C]α-MDG in control mice decreased in the presence of unlabelled α-MDG, whereas those in mutant mice were not affected in the presence of unlabelled α-MDG (Fig. 2e, right panel, and Supplementary Fig. 11). Thus, the uptake system for α-MDG was almost completely abolished in the mutant mice at age three weeks. Similar experiments were also performed to investigate the fundamental roles of Rab8 in the functional expression of oligopeptide transport system(s) (Fig. 2f and Supplementary Fig. 11). The saturable uptake system for Gly-Sar was also almost completely abolished in the mutant mice at age three weeks. These findings suggest that the mutant mice were suffering from starvation because nutrient uptake was almost abolished in the small intestine. In support of this, the level of the starvation marker 3-hydroxy-butyrate in the blood was elevated in the mutant mice (Fig. 2d)

To determine whether the above-mentioned abnormalities in the small intestine are the major cause of death, we generated small-intestine-specific knockout by breeding *Rab8*$^{flox/+}$ mice with transgenic mice expressing Cre recombinase under the control of the villin (*Vil1*) promoter (*Vil1-cre*)[18]. All the small intestine-specific knockout mice that had only one floxed allele (*Vil1-cre*; *Rab8*$^{flox/-}$) died with a similar time course to conventional knockout mice (Supplementary Fig. 4e), whereas those with two floxed alleles (*Vil1-cre*; *Rab8*$^{flox/flox}$) were relatively healthy at earlier postnatal ages but died within approximately 12 weeks of birth, probably because the efficiency of Rab8 deletion is reduced in the latter genotype. When we observed intestinal epithelial cells from healthy conditional knockout mice (*Vil1-cre*; *Rab8*$^{flox/flox}$), which showed mosaic expression of Rab8, Rab8-negative cells had large sub-apical vacuoles identical to those in the knockout mice, whereas neighbouring Rab8-positive cells had no such vacuoles (Fig. 1d). This result demonstrates that (1) the malfunction of the small intestine is responsible for the death of the mutant mice, (2) the abnormal sub-apical inclusion and shortening of microvilli result primarily from Rab8 deficiency and (3) the effect of *Rab8* is cell-autonomous. In support of these data, Rab8 is most highly expressed in the small intestine (Supplementary Fig. 4d),

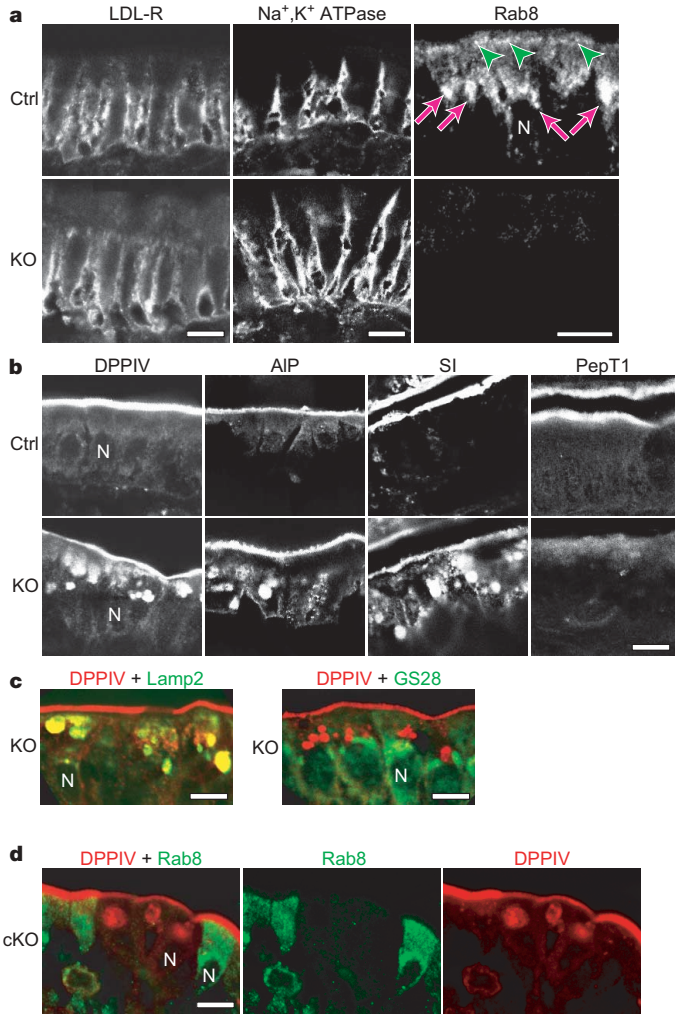


Figure 1 | Accumulation of apical but not basolateral proteins in cytoplasm of Rab8-deficient intestinal epithelial cells. **a**, Localization of basolateral markers, LDL-R (left panels) and Na^+,K^+ ATPase (middle panels), and Rab8 (right panels); KO, *Rab8*$^{-/-}$; Ctrl, *Rab8*$^{+/+}$ or *Rab8*$^{+/-}$; N, nucleus. **b**, Localization of apical markers (DPPIV, AlP, SI, and PepT1). **c**, Left panel, colocalization of DPPIV (red) and Lamp2 (green); right panel, distinct localization of DPPIV (red) and the Golgi marker GS28 (green). **d**, A sample with mosaic Rab8 expression selected from small-intestine-specific knockout mice. DPPIV (red) is localized intracellularly in only Rab8-negative enterocytes. cKO, conditional knockout; Ctrl, control. Separate images for **c** are shown in Supplementary Fig. 6. Scale bars, 10 μm.

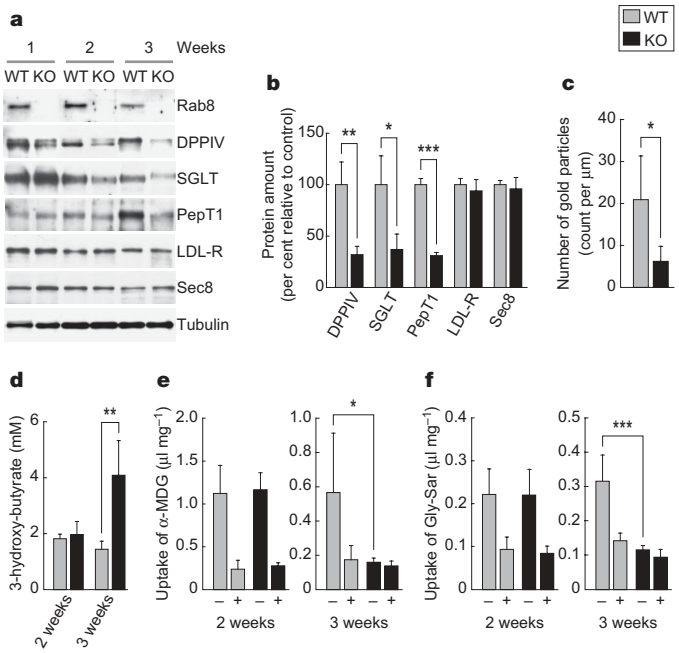


Figure 2 | The reduction of apical markers in the small intestine of the knockout mice causes a reduction of substrate uptake from the apical side. **a**, Immunoblot analysis of crude extracts of small intestine from wild-type (WT) and knockout (KO) mice. **b**, Quantification of western blot products from 3-week-old mouse intestinal samples. Data are normalized against data of the wild-type mice. Values represent mean ± s.d. of three mice. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ (Student's *t*-test). **c**, Density of gold particles labelled against DPPIV per apical width of 3-week-old-mouse intestinal epithelial cells. Values represent mean ± s.d. ($n = 8$ cells in wild-type and $n = 18$ cells in knockout). **d**, Serum levels of 3-hydroxy-butyrate in wild-type and knockout mice. Values represent mean ± s.d. of 3–8 mice. **e**, **f**, Uptake rates of α-MDG (**e**) and Gly-Sar (**f**) from the apical side of small intestines of wild-type and knockout mice. Small intestines were isolated from 2-week-old mice (left panel) and 3-week-old mice (right panel), and uptake of each labelled compound was assessed in the absence (−) or presence (+) of unlabelled compound. Data were normalized by medium concentration and represent mean ± s.d. of 3–8 mice. Time profiles and uptake rates of extracellular markers in identical samples are shown in Supplementary Fig. 11.

and its expression level is highest in 2-week-old wild-type mice (Fig. 2a).

When we examined the small intestine by electron microscopy, a marked shortening of microvilli was observed in the epithelial cell of the mutant mice at age three weeks (Fig. 3a–c). However, the lengths of microvilli were similar between the mutant and control mice at age two weeks (Fig. 3c). In addition, we observed a markedly increased number of enlarged organelles with electron-dense materials within them in only the mutant intestinal epithelial cells (Fig. 3e, f). These organelles were strongly positive for DPPIV as revealed by immuno-electron-microscopy (Supplementary Fig. 12). The density of gold particles on the apical plasma membrane was markedly reduced (30% of that of the wild-type cells) in 3-week-old knockout enterocytes (Fig. 2c). Taken together this suggests that, in Rab8-knockout mice, apically destined peptidases and transporters are mislocalized to lysosomes and then degraded, which causes reduction both in terminal digestion and absorption from the apical plasma membrane, ultimately leading to malnutrition and the death of the mice (Supplementary Fig. 1). The reason for this relatively tissue-specific phenotype is not completely clear but can be explained in part by the expression pattern and the pathway of apical transport that the tissues use. Namely, (1) the small intestine is one of the tissues that contain the highest expression level of Rab8 and the onset of the phenotype coincides with the peak of Rab8 expression (that is, two weeks postnatal), and (2) tissues that predominantly use transcytosis for apical localization of proteins are less likely to be affected, considering the normal phenotype of the liver.

Interestingly, we frequently observed vacuoles with strongly positive F-actin staining at their periphery only in the mutant intestinal cells (Fig. 3g). By electron microscopy, we observed vacuoles with microvilli radiating towards their lumens (Fig. 3h, i). These vacuoles are identical in shape with microvillus inclusion bodies observed in the small intestine of patients with the human hereditary disease, microvillus inclusion disease[19,20]. To determine whether these vacuoles are produced by endocytosis, we fed the knockout mice fluorescent markers for two days and stained the F-actin of the small intestine. We observed that the fluorescent markers were internalized within F-actin-encircled vacuoles as well as within a small number of DPPIV-positive lysosomes (Fig. 3j and Supplementary Fig. 13).

Because Rab8 knockout mice show almost identical phenotypes (for example, diarrhoea, malnutrition, shortening of microvilli, and microvillus inclusion) (Fig. 3) to patients with human microvillus inclusion disease, we sequenced the genome of the patients. However, we identified no mutations in the exons of *RAB8* genes of the patients of either early-onset (two cases) or late-onset (one case) form. However, one of the patients showed a marked decrease in the level of RAB8 protein in the small intestine by immunohistochemistry (Fig. 4). In the small intestine of this patient, *RAB8* messenger RNA was reduced, particularly in the enterocyte-enriched fraction as shown by quantitative reverse transcriptase (RT)–PCR (Supplementary Fig. 14), suggesting that some defects in transcriptional regulation of RAB8 may be linked to the pathogenesis of this patient. This finding indicates that the phenotype of microvillus

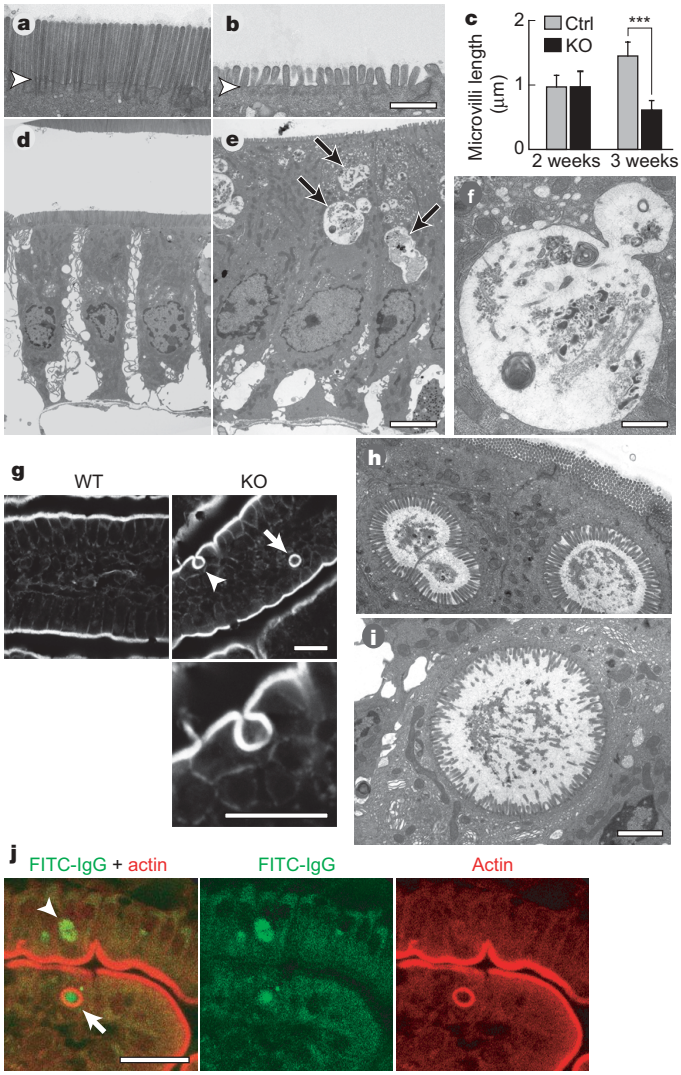


Figure 3 | Microvillus atrophy and inclusions in Rab8-deficient mice. **a**, **b**, **d**–**f**, Electron micrographs of 3-week-old-mouse intestinal cells from the knockout (**b**, **e**) and the wild-type mice (**a**, **d**). Intracellular vacuoles are shown in **e** (arrows) and **f**. **c**, Graph showing the average microvillus length of epithelial cells. Values represent mean ± s.d. ($n > 23$ cells from three mice per group. $^{***}P < 0.001$; Student's *t*-test). **g**–**i**, Microvillus inclusions in 3-week-old knockout intestinal epithelial cells. **g**, Phalloidin-stained cells from the wild-type (upper left panel) and the knockout (upper right and lower right panels) mice. An intracellular spheroid (arrow, upper right panel) and an invagination of apical plasma membranes (arrowhead, upper right panel, and lower right panel) are shown. **h**, **i**, Electron micrographs of microvillus inclusions in the knockout mice. **j**, A fluid-phase marker, fluorescein isothiocyanate-IgG (FITC-IgG) is endocytosed in a strongly F-actin-positive vacuole (left panel, arrow) as well as in F-actin-negative structures (left panel, arrowhead). Scale bars, 1 μm (**a**, **b**, **f**), 5 μm (**d**, **e**), 20 μm (**g**, **j**) and 2 μm (**h**, **i**).

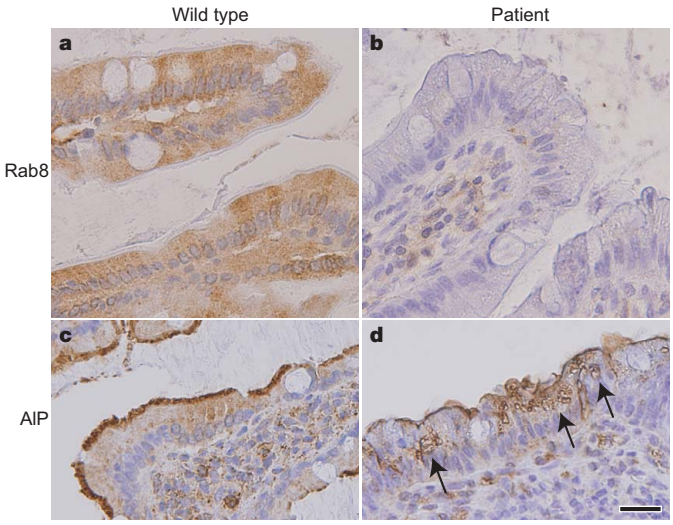


Figure 4 | RAB8 is greatly reduced in intestinal epithelial cells of the microvillus inclusion disease patient. **a**, **b**, Abundant expression of RAB8 in the control human epithelial cells (**a**) and markedly reduced expression in the patient's epithelial cells (**b**) are shown by immunohistochemistry. **c**, **d**, The apical marker alkaline phosphatase (**c**) is localized intracellularly (arrows) in the patient's intestinal epithelial cells (**d**). Scale bar, 50 μm.

inclusion disease could be rescued by re-expression of the *RAB8* gene because reversion of only one allele of the *RAB8* gene in the reverted mice ($RAB8^{flox/-}$) led to restoration of microvilli and other ultrastructures (Supplementary Fig. 5). However, because there are at least three clinical groups of microvillus inclusion disease (early-onset, late-onset and atypical forms), a number of genes in addition to *RAB8* are likely to be involved in a common trafficking pathway, in which defects lead to this disease.

METHODS

Generation of Rab8-deficient mice. *Rab8* knockout mice were generated essentially as previously described[21] except that a *Bcl2* splice-acceptor–internal ribosomal entry site (SA–IRES) cassette[11] fused with a promoterless beta-geo cassette was used for constructing the targeting vector. Additional construct information is available in Supplementary Methods. To generate small-intestine-specific knockout mice, we crossed $Rab8^{geo/+}$ mice with *Act-flp-e* transgenic mice (Jackson Laboratory)[10] and obtained $Rab8^{flox/+}$ mice. We then crossed $Rab8^{flox/+}$ mice with *Vil1-cre* transgenic mice (Jackson Laboratory). To generate nullizygous mice ($Rab8^{-/-}$), we crossed $Rab8^{flox/+}$ mice with *CAG-cre* transgenic mice[22] or *CMV-cre* transgenic mice (Jackson Laboratory)[23]. An allele flanked by two *loxP* sites is a 'flox' allele; an allele that has a β-galactosidase and neomycin resistance gene (β-geo) is called a 'geo' allele.

Histology and western blot analysis. An antibody against Rab8 (peptide C-KAKMDKKLEGNSPQGSNQGVK) was raised in rabbits and affinity-purified. Immunofluorescence microscopy, immunoelectron microscopy and western blot analysis were performed as previously described[21,24,25]. For western blot analysis, 10 μg of protein was loaded per lane.

Uptake studies in everted intestinal sac. An everted intestinal sac was prepared from the upper part (within 10-cm below the pylorus) of the small intestines as described previously[17].

Analysis of transgenic *C. elegans*. Live worms expressing *GFP–rab-8* and *GFP–pgp-1* were generated as previously described[26]. Other worms were obtained and observed by a confocal microscope as described in Supplementary Methods.

Analysis of human samples. We obtained DNA samples from the blood of individuals diagnosed as having microvillus inclusion disease by an examination of small-intestine biopsy samples carried out by medical pathologists. Small-intestine samples were taken from the fixed and non-fixed samples preserved after the removal of the intestine for transplantation. The patient discussed in this paper has already been described[27,28].

Please refer to Supplementary Information for additional Methods details.

Received 15 April; accepted 15 May 2007.
Published online 27 June 2007.

1. Nelson, W. J. Adaptation of core mechanisms to generate cell polarity. *Nature* **422,** 766–774 (2003).
2. Schuck, S. & Simons, K. Polarized sorting in epithelial cells: raft clustering and the biogenesis of the apical membrane. *J. Cell Sci.* **117,** 5955–5964 (2004).
3. Rodriguez-Boulan, E., Musch, A. & Le Bivic, A. Epithelial trafficking: new routes to familiar places. *Curr. Opin. Cell Biol.* **16,** 436–442 (2004).
4. Matter, K. & Mellman, I. Mechanisms of cell polarity: sorting and transport in epithelial cells. *Curr. Opin. Cell Biol.* **6,** 545–554 (1994).
5. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* **2,** 107–117 (2001).
6. Sabatini, D. D. In awe of subcellular complexity: 50 years of trespassing boundaries within the cell. *Annu. Rev. Cell Dev. Biol.* **21,** 1–33 (2005).
7. Louvard, D., Kedinger, M. & Hauri, H. P. The differentiating intestinal epithelial cell: establishment and maintenance of functions through interactions between cellular structures. *Annu. Rev. Cell Biol.* **8,** 157–195 (1992).
8. Huber, L. A. *et al.* Rab8, a small GTPase involved in vesicular traffic between the TGN and the basolateral plasma membrane. *J. Cell Biol.* **123,** 35–45 (1993).
9. Ang, A. L. *et al.* The Rab8 GTPase selectively regulates AP-1B-dependent basolateral transport in polarized Madin-Darby canine kidney cells. *J. Cell Biol.* **163,** 339–350 (2003).
10. Rodriguez, C. I. *et al.* High-efficiency deleter mice show that FLPe is an alternative to Cre-*loxP*. *Nature Genet.* **25,** 139–140 (2000).
11. Ishida, Y. & Leder, P. RET: a poly A-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. *Nucleic Acids Res.* **27,** e35 (1999).
12. Homareda, H., Nagano, Y. & Matsui, H. Immunochemical identification of exposed regions of the Na^+,K^+-ATPase α-subunit. *FEBS Lett.* **327,** 99–102 (1993).
13. Umesaki, Y., Tohyama, K. & Mutai, M. Biosynthesis of microvillus membrane-associated glycoproteins of small intestinal epithelial cells in germ-free and conventionalized mice. *J. Biochem.* **92,** 373–379 (1982).
14. Saito, H. *et al.* Cloning and characterization of a rat H+/peptide cotransporter mediating absorption of beta-lactam antibiotics in the intestine and kidney. *J. Pharmacol. Exp. Ther.* **275,** 1631–1637 (1995).
15. Koike, M. *et al.* Participation of autophagy in storage of lysosomes in neurons from mouse models of neuronal ceroid-lipofuscinoses (Batten disease). *Am. J. Pathol.* **167,** 1713–1728 (2005).
16. Nakamura, N. *et al.* Association of mouse sorting nexin 1 with early endosomes. *J. Biochem.* **130,** 765–771 (2001).
17. Tamai, I. *et al.* The predominant contribution of oligopeptide transporter PepT1 to intestinal absorption of beta-lactam antibiotics in the rat small intestine. *J. Pharm. Pharmacol.* **49,** 796–801 (1997).
18. Madison, B. B. *et al. cis* elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277,** 33275–33283 (2002).
19. Davidson, G. P., Cutz, E., Hamilton, J. R. & Gall, D. G. Familial enteropathy: a syndrome of protracted diarrhea from birth, failure to thrive, and hypoplastic villus atrophy. *Gastroenterology* **75,** 783–790 (1978).
20. Phillips, A. D., Jenkins, P., Raafat, F. & Walker-Smith, J. A. Congenital microvillous atrophy: specific diagnostic features. *Arch. Dis. Child.* **60,** 135–140 (1985).
21. Harada, A. *et al.* Altered microtubule organization in small-calibre axons of mice lacking tau protein. *Nature* **369,** 488–491 (1994).
22. Sakai, K. & Miyazaki, J. A transgenic mouse line that retains Cre recombinase activity in mature oocytes irrespective of the cre transgene transmission. *Biochem. Biophys. Res. Commun.* **237,** 318–324 (1997).
23. Schwenk, F., Baron, U. & Rajewsky, K. A *cre*-transgenic mouse strain for the ubiquitous deletion of *loxP*-flanked gene segments including deletion in germ cells. *Nucleic Acids Res.* **23,** 5080–5081 (1995).
24. Harada, A., Sobue, K. & Hirokawa, N. Developmental changes of synapsin I subcellular localization in rat cerebellar neurons. *Cell Struct. Funct.* **15,** 329–342 (1990).
25. Harada, A. *et al.* MAP2 is required for dendrite elongation, PKA anchoring in dendrites, and proper PKA signal transduction. *J. Cell Biol.* **158,** 541–549 (2002).
26. Sato, M. *et al. Caenorhabditis elegans* RME-6 is a novel regulator of RAB-5 at the clathrin-coated pit. *Nature Cell Biol.* **7,** 559–569 (2005).
27. Kagitani, K. *et al.* Hypophosphatemic rickets accompanying congenital microvillous atrophy. *J. Bone Miner. Res.* **13,** 1946–1952 (1998).
28. Sasaki, T. *et al.* Zoom endoscopic evaluation of rejection in living-related small bowel transplantation. *Transplantation* **73,** 560–564 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Takano, T. Horie, R. Hirai, Y. Okada, C. Ohsawa, M. Sugiura, H. Hata and R. Ishida for assistance with cell culture, uptake studies, animal care, and microscopy, H. Gomi, T. Izumi, M. Komachi, C. Mogi, H. Tomura, F. Okajima, N. Shihara and T. Suzuki for teaching us various techniques, Y. Wada, Y. Uchiyama, H. Homareda, K. Inui and Y. Umesaki for providing antibodies, J. Miyazaki for providing the transgenic mice, S. Mitani for providing rab-8 knockout worms, and Y. Ishida for providing the construct. We thank the affected individuals and their families for providing the blood samples. The Lamp2 monoclonal antibody was obtained from the Developmental Studies Hybridoma Bank developed under the auspices of the NICHD and maintained by The University of Iowa. We also thank H.-P. Zimmer, A. Ballauff and T. Berger for providing DNA samples of the patients. This work was supported by grants-in-aid and the 21st century Center of Excellence Program from the Japanese Ministry of Education, Culture, Sports, Science and Technology to T.S., M.S., K.S. and A.H.

Author Contributions T.S. generated and analysed Rab8 knockout mice. S.M. took care of the microvillus inclusion disease patient and collected the blood and small intestine samples. Y. Kato supervised the measurement of the absorption rate of nutrients in the intestine of the mice. K.S. and M.S. generated the transgenic *C. elegans* and photographed them. N.T. generated the chimeric mice. K.M. took care of the microvillus inclusion disease patient. K.O. supervised the treatment of the patient. Y. Kubo measured the absorption rate of nutrients in the intestine. A.T. provided the instruments and supervised the absorption experiments. R.H. stained the tissue, arranged figures and wrote part of the manuscript. A.H. planned and supervised the experiments, performed morphological analyses, arranged figures and wrote the manuscript

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.H. (aharada@showa.gunma-u.ac.jp).

LETTERS

A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity

Tracy R. Rosebrock[1,2], Lirong Zeng[1], Jennifer J. Brady[1], Robert B. Abramovitch[1,2], Fangming Xiao[1] & Gregory B. Martin[1,2]

Many bacterial pathogens of plants and animals use a type III secretion system to deliver diverse virulence-associated 'effector' proteins into the host cell[1]. The mechanisms by which these effectors act are mostly unknown; however, they often promote disease by suppressing host immunity[2]. One type III effector, AvrPtoB, expressed by the plant pathogen *Pseudomonas syringae* pv. *tomato*, has a carboxy-terminal domain that is an E3 ubiquitin ligase[3]. Deletion of this domain allows an amino-terminal region of AvrPtoB (AvrPtoB$_{1–387}$) to be detected by certain tomato varieties leading to immunity-associated programmed cell death[4]. Here we show that a host kinase, Fen, physically interacts with AvrPtoB$_{1–387}$ and is responsible for activating the plant immune response. The AvrPtoB E3 ligase specifically ubiquitinates Fen and promotes its degradation in a proteasome-dependent manner. This degradation leads to disease susceptibility in Fen-expressing tomato lines. Various wild species of tomato were found to exhibit immunity in response to AvrPtoB$_{1–387}$ and not to full-length AvrPtoB. Thus, by acquiring an E3 ligase domain, AvrPtoB has thwarted a highly conserved host resistance mechanism.

Pseudomonas syringae pv. *tomato* (*Pst*) causes bacterial speck disease of tomato (*Solanum lycopersicum*) by using its type III secretion system to deliver about 30 effectors into the plant cell[5]. Tomato varieties that are immune to speck disease express the Pto kinase that detects either of two *Pst* effector proteins, AvrPto or AvrPtoB. This detection involves the physical interaction of Pto with AvrPto or AvrPtoB and results in the rapid activation of an array of host defence responses including localized programmed cell death (PCD)[6]. The *Pto* gene was isolated from the wild tomato species *S. pimpinellifolium* and is a member of a small clustered gene family[6,7] (Fig. 1a). Embedded within this gene cluster is the *Prf* gene that encodes a leucine-rich repeat-containing protein that physically interacts with Pto and is required for Pto-mediated immunity[6,8]. Another member of the *Pto* gene family encodes the Fen kinase, which shares 80% amino acid identity with Pto but does not recognize AvrPto or AvrPtoB[9,10].

AvrPtoB is a modular protein[4]. An N-terminal region (AvrPtoB$_{1–307}$) is sufficient to elicit Pto/Prf-mediated PCD, whereas a C-terminal region (AvrPtoB$_{388–553}$) contains a domain that is an E3 ubiquitin ligase[3,11]. Two truncations of AvrPtoB (AvrPtoB$_{1–509}$ and AvrPtoB$_{1–387}$) lacking the E3 ligase domain, or AvrPtoB point mutants compromised in E3 ligase activity, are detected by tomato varieties lacking the Pto kinase and by a tobacco species, *Nicotiana benthamiana*[3,4]. This Pto-independent detection leads to PCD, is dependent on Prf and is referred to as Rsb (for 'resistance suppressed by AvrPtoB C terminus')[4]. AvrPtoB$_{1–307}$ does not elicit Rsb immunity, indicating that a domain between amino acid residues 308 and 387 is required for Rsb-associated recognition (Supplementary Fig. 1). Here we identify the host protein responsible for Rsb and explain the mechanism by which AvrPtoB E3 ligase activity suppresses this recognition.

Dependence on Prf indicated that Rsb might involve a member of the Pto family. Four members of the *Pto* gene family in *S. pimpinellifolium* are transcribed in leaves (*Fen*, *PtoC*, *PtoD* and *Pto*), of which only two, *Pto* and *Fen*, encode active kinases[12]. The kinase activity of Pto is necessary for immunity, leading us to propose that the protein responsible for Rsb would also be an active kinase[6]. To determine whether a *Pto* family member is involved in Rsb immunity, we examined a tomato line, RG-PtoR(hpPto), that is a stable transformant knocked down for expression of the *Pto* gene family by RNA-mediated interference[13]. RG-PtoR(hpPto) plants were inoculated with *Pst* DC3000 strains delivering either AvrPtoB or AvrPtoB$_{1–509}$ (Fig. 1b). In leaves of RG-PtoR(hpPto) plants, both of these strains reached populations similar to that of the control susceptible line, RG-prf3, indicating a complete loss of Pto-mediated immunity and the Rsb phenotype. As expected, RG-PtoR plants were resistant to both *Pst* strains, whereas RG-pto11, which lacks Pto, showed it had the Rsb phenotype by being resistant to only the strain delivering AvrPtoB$_{1–509}$. By using virus-induced gene silencing in *N. benthamiana*, we found that knocked-down expression of the *Pto* gene family, *Prf*, or previously described components of the Pto-mediated PCD-associated signalling pathway compromised the Rsb phenotype[14] (Fig. 1c, and Supplementary Fig. 2). These results implicated one or more members of the *Pto* family in Rsb immunity.

There is a strict correlation between immunity conferred by Pto and the ability of this kinase to interact with AvrPtoB in the yeast two-hybrid system[10]. We proposed that the host protein responsible for Rsb would interact with AvrPtoB$_{1–387}$ and not with the non-Rsb-eliciting fragment, AvrPtoB$_{1–307}$. Using the yeast two-hybrid system we tested the Pto family members for interaction with AvrPtoB, AvrPtoB$_{1–387}$ or AvrPtoB$_{1–307}$ (Fig. 1d). Of these, only Fen interacted exclusively with AvrPtoB$_{1–387}$. PtoC interacted with AvrPtoB$_{1–387}$ but also with AvrPtoB$_{1–307}$. Pto, as expected, interacted with AvrPtoB and both AvrPtoB truncations, whereas PtoA and PtoD showed no interactions. Western blots confirmed the expression of each of the Pto family proteins and AvrPtoB proteins (Fig. 1d, and Supplementary Fig. 3).

To determine whether Fen or PtoC activates immunity, we expressed each of these proteins with AvrPtoB, AvrPtoB$_{1–387}$ or AvrPtoB$_{1–307}$ in protoplasts of tomato RG-PtoS or *N. benthamiana* (Fig. 1e, Supplementary Fig. 4A). Fen activated PCD when expressed with AvrPtoB$_{1–387}$ but not with AvrPtoB or AvrPtoB$_{1–307}$, whereas PtoC was unable to initiate PCD when expressed with any of these AvrPtoB proteins. Western blots confirmed expression of the Pto family proteins (Fig. 1e, and Supplementary Fig. 4B). As expected, Pto activated PCD in tomato when expressed with AvrPtoB, AvrPtoB$_{1–387}$ or AvrPtoB$_{1–307}$. The inability of PtoC to activate

[1]Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, New York 14853, USA. [2]Department of Plant Pathology, Cornell University, Ithaca, New York 14853, USA.

effector-elicited PCD, its interaction with the non-Rsb eliciting fragment AvrPtoB$_{1–307}$ and its lack of kinase activity[12] excluded a function for this protein in Rsb. Taken together, these data indicate that Fen is responsible for Rsb immunity.

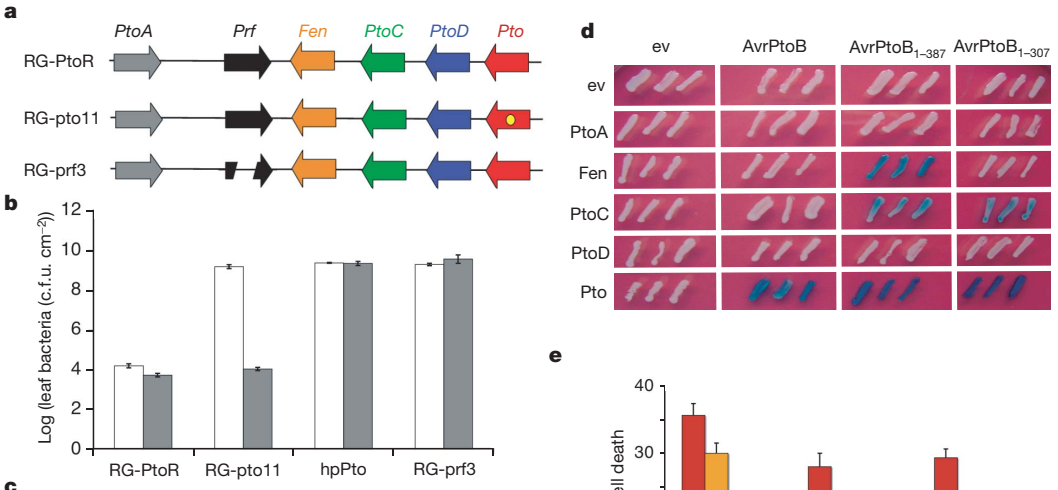
It was possible that the lack of interaction between Fen and AvrPtoB in yeast (Fig. 1d) involved the E3 ligase activity of AvrPtoB. We therefore tested whether Fen interacted with E3 ligase-deficient AvrPtoB mutants. The AvrPtoB-Quad protein (Quad) contains four mutations in critical lysine residues[11], and a second mutant, E2BS, has three point mutations at predicted E2-binding sites[3]. Both AvrPtoB mutants interacted with Fen and PtoC, indicating that the E3 ligase activity interferes with certain protein interactions (Fig. 2a). As expected, the altered proteins interacted with Pto and were unable to interact with PtoA or PtoD.

We proposed that the AvrPtoB E3 ligase might ubiquitinate Fen to disrupt recognition of the AvrPtoB N-terminal region. To test this possibility, ubiquitination assays were performed *in vitro* with Fen and AvrPtoB and a series of controls (Fig. 2b). Among five kinases tested, only Fen was ubiquitinated in the presence of AvrPtoB, as indicated by the appearance of high-molecular-mass Fen proteins (Fig. 2b). Absence of E1, E2 or AvrPtoB proteins in the assay abolished the high-molecular-mass forms of Fen (Fig. 2b). In particular, Pto was not ubiquitinated by AvrPtoB in this assay. Fen ubiquitination might be due to unique ubiquitination sites (namely lysine residues) in this kinase. There are only five lysine residues that are present in Fen but absent from Pto and PtoC (Lys 70, Lys 72, Lys 155, Lys 253 and Lys 290; Supplementary Fig. 5A). Of these, Lys 70 and Lys 72 are close to the ATP-binding site (Lys 69) of Fen, raising the possibility that this region might be targeted for ubiquitination. However, arginine substitutions at any one of the five lysine residues had no effect on Fen ubiquitination (Supplementary Fig. 5B), suggesting that either multiple lysine residues are ubiquitinated or other structural differences between Pto and Fen account for the differential ubiquitination.

Ubiquitination of Fen by AvrPtoB raised the possibility that, in the plant cell, AvrPtoB E3 ligase activity might target Fen for degradation. To test this, we expressed Fen, PtoC or Pto with AvrPtoB or Quad proteins in RG-prf3 tomato protoplasts and assessed protein abundance. Fen accumulated poorly in the presence of AvrPtoB, reaching only about 35% of the abundance with the Quad protein (Fig. 3a). In contrast, Pto and PtoC accumulated in the presence of AvrPtoB to levels comparable to their abundance with the Quad protein (Fig. 3a).

Ubiquitination often marks a protein for degradation by means of the 26S proteasome. If AvrPtoB targets Fen for degradation, then inhibition of the proteasome should allow Fen to accumulate in the presence of AvrPtoB. In RG-prf3 protoplasts, we expressed Fen or Pto with AvrPtoB or Quad in the presence of MG132, a proteasome inhibitor. Treatment with MG132 resulted in a roughly 80% increase in Fen accumulation when expressed with AvrPtoB (Fig. 3b). MG132 had no effect on Fen coexpressed with Quad or with Pto coexpressed with AvrPtoB or Quad. Similar results were seen with a second proteasome inhibitor, MG115 (data not shown). We tested the effect of a general plant protease inhibitor cocktail on Fen abundance and found no effect on Fen accumulation (Fig. 3c). These results support a function for the AvrPtoB E3 ligase in proteasome-dependent degradation of the Fen kinase.

The specific targeting of Fen by AvrPtoB to suppress Rsb immunity suggested a selective advantage for the N-terminal region of AvrPtoB to acquire and maintain the E3 ligase domain. We therefore examined whether Rsb immunity is conserved among cultivated and wild species of tomato. Cultivars Moneymaker (MM), VFNT Cherry



c

Vector	Pto/AvrPto	AvrPtoB$_{1–387}$	MAPKKKα	Bax
TRV	PCD	PCD	PCD	PCD
TRV:*Pto*	–	–	PCD	PCD
TRV:*Prf*	–	–	PCD	PCD
TRV:*MAPKKKα*	–	–	PCD	PCD
TRV:*MEK2*	–	–	–	PCD
TRV:*MEK1*	–	–	–	–
TRV:*SIPK*	–	–	–	PCD
TRV:*WIPK*	–	–	–	PCD

Figure 1 | The Fen kinase is responsible for the Rsb phenotype. a, Genome organization of the *Pto* family and *Prf* from *Solanum pimpinellifolium*. Rio Grande-PtoR (RG-PtoR) plants are wild type[6], RG-pto11 plants have a deleterious point mutation (yellow dot) in *Pto*, and RG-prf3 plants have a deletion in *Prf*[6]. **b**, Growth of *Pst* strains delivering AvrPtoB (white) or AvrPtoB$_{1–509}$ (grey) in RG-PtoR, RG-pto11, RG-prf3 and RG-PtoR(hpPto) leaves. Error bars represent s.e.m. ($n = 6$). c.f.u., colony-forming units. **c**, Response after transient expression of Pto with AvrPto, AvrPtoB$_{1–387}$, mitogen-activated protein kinase kinase kinase (MAPKKK)α or Bax in *N. benthamiana* leaves silenced for genes in the Pto pathway[14] (listed at the left) (see Supplementary Fig. 2). MEK, MAP kinase/ERK kinase; SIPK, salicylic-acid-induced protein kinase; WIPK, wound-induced protein kinase; dashes indicate no PCD. **d**, Pto family members or empty vector (ev) tested for interaction with forms of AvrPtoB in the yeast two-hybrid system. Blue patches show positive interactions. **e**, Response of RG-PtoS protoplasts after coexpression of AvrPtoB$_{1–387}$ with Pto (red columns), Fen (orange columns), PtoC (green columns) or empty vector (white columns). Data are presented as percentage cell death in experimental samples after subtraction of cell death percentages occurring with empty vector. Error bars represent s.e.m. ($n = 3$).

(VFNT-C), and Ailsa Craig (AC) were tested for the presence of Rsb by infiltrating their leaves with two Rsb-eliciting *Pst* strains (DC3000Δ*avrPtoB* plus AvrPtoB$_{1–387}$, or DC3000Δ*avrPtoB* plus Quad) or two non-Rsb-eliciting strains (DC3000Δ*avrPtoB* plus AvrPtoB, or DC3000Δ*avrPtoB* plus AvrPtoB$_{1–307}$). RG-PtoR, RG-pto11 and RG-prf3 lines were included as controls. Rsb immunity was activated in MM, VFNT-C and AC and, as expected, in RG-pto11 (Fig. 4a).

To test whether Rsb immunity was activated by a *S. lycopersicum* orthologue of Fen, we examined a previously described Fen allele from AC (AC-fen)[15]. As assessed by ion leakage, expression in *N. benthamiana* leaves of AC-fen with AvrPtoB$_{1–387}$, but not with AvrPtoB$_{1–307}$, caused PCD comparable to the expression of Pto or Fen with AvrPtoB$_{1–387,}$ (Fig. 4b). As expected, only Pto initiated PCD when expressed with AvrPtoB$_{1–307}$. Orthologues of Fen can therefore confer Rsb immunity. A screen of 37 accessions of wild species of tomato revealed that 21 express Rsb immunity but only 5 seem to have Pto-mediated immunity (Supplementary Table 1; Fig. 4c shows three examples). The wide occurrence of Rsb is consistent with phylogenetic analyses indicating that the *Fen* gene arose before the *Pto* gene in the *Solanum* species[16] and suggests that the E3 ligase activity of AvrPtoB evolved to circumvent a broadly conserved host resistance mechanism.

The interaction of AvrPtoB$_{1–387}$ with Fen kinases might indicate that the effector evolved to interfere with a possible function of these kinases in basal defence. In fact, E3 ligase-deficient AvrPtoB proteins are able to suppress certain basal defences[17,18]. We therefore compared the growth of a type III secretion-defective *Pst* strain (DC3000Δ*hrcQ-U* (ref. 19)) in wild-type tomato plants and in a line overexpressing Fen (Supplementary Fig. 6). Similar overexpression experiments have demonstrated a function for the *Arabidopsis* protein RIN4 as a negative regulator of basal defence[20]. However, we found no effect of Fen overexpression on growth of the DC3000Δ*hrcQ-U* strain (Supplementary Fig. 6).

Molecular mimicry of host proteins by bacterial pathogens is common; however, so far only a handful of bacterial proteins are known to manipulate the host ubiquitination system[3,21–23]. Of these, AvrPtoB, *Salmonella* SopA, the *Shigella* IpaH family and *Salmonella* SspH1 have intrinsic E3 ubiquitin ligase activity, potentially releasing them from dependence on host proteins to secure substrate specificity[22,23]. Interestingly, IpaH9.8 and SspH1 also ubiquitinate protein kinases, although the relevance of this activity to the host–pathogen interaction has not been established[24]. Overexpression of the AvrPtoB E3 ligase in leaves suppresses PCD induced by the pro-apoptotic protein Bax and by many, but not all, plant resistance proteins[4,17,18,25]. AvrPtoB also suppresses PCD in yeast[4]. Our present data do not explain these general PCD suppression activities; they may be due to a second, highly conserved target of AvrPtoB or to

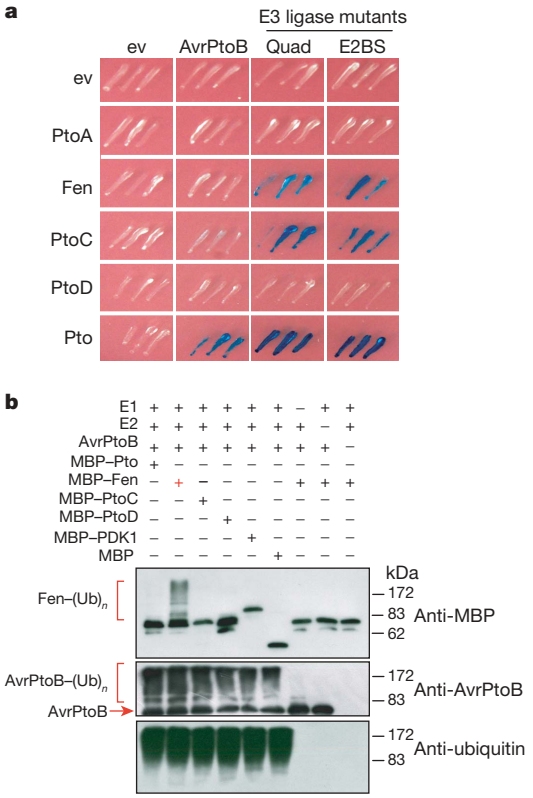


Figure 2 | Fen is ubiquitinated in the presence of AvrPtoB. **a**, Pto family members or empty vector (ev) tested for interaction with AvrPtoB, or AvrPtoB proteins with compromised E3 ligase activity (Quad and E2BS) in the yeast two-hybrid system. Blue patches show positive interactions. **b**, Immunoblots of an *in vitro* ubiquitination assay with recombinant E1, E2, AvrPtoB, ubiquitin (Ub) and MBP fusions of the kinases shown. Polyubiquitination of Fen is indicated by (Fen–(Ub)$_n$). Polyubiquitinated forms of AvrPtoB (AvrPtoB–(Ub)$_n$) confirmed E3 ligase activity. An anti-Ub antibody detected the incorporation of ubiquitin in each assay.

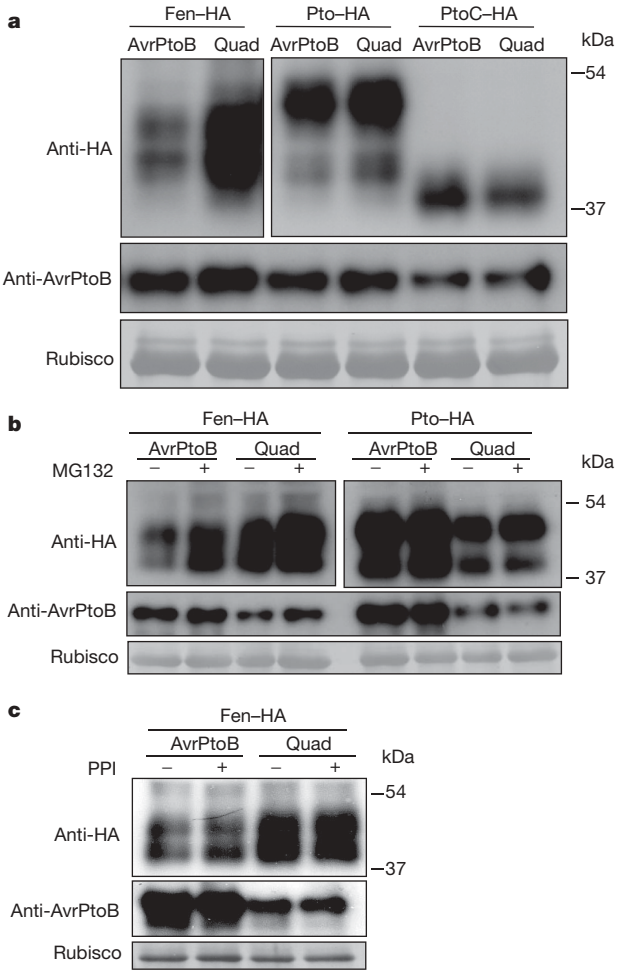


Figure 3 | In the presence of AvrPtoB, Fen is degraded in a proteasome-dependent manner. **a**, Immunoblot to detect accumulation of Fen, Pto or PtoC in RG-prf3 protoplasts in the presence of AvrPtoB or Quad. Visualization of ribulose 1,5-bisphosphate carboxylase–oxygenase (Rubisco) subunits by Coomassie blue confirmed equal loading. The Fen panel shows a roughly 20-s exposure, in contrast with about 10 s for the others; however, the images are from the same experiment. HA, haemagglutinin. **b**, As in **a** after treatment with MG132 (minus, dimethylsulphoxide (DMSO) only; plus, with 50 μM MG132). **c**, As in **a** with the addition of general plant protease inhibitor (PPI) cocktail (minus, DMSO only; plus, 0.2% PPI).

overexpression of the effector. However, our data fully account for the immunity suppression observed with AvrPtoB-expressing *Pseudomonas* strains on infection of tomato leaves lacking Pto[3,4,11]. Furthermore, our results suggest that Pto evolved not only to recognize AvrPto and AvrPtoB but also to be invulnerable to AvrPtoB-mediated ubiquitination and subsequent degradation.

a

Line	AvrPtoB	AvrPtoB$_{1–307}$	AvrPtoB$_{1–387}$	Quad
RG-PtoR	R	R	R	R
RG-pto11	S	S	R	R
RG-prf3	S	S	S	S
MM	S	S	R	R
VFNT-C	S	S	R	R
AC	S	S	R	R

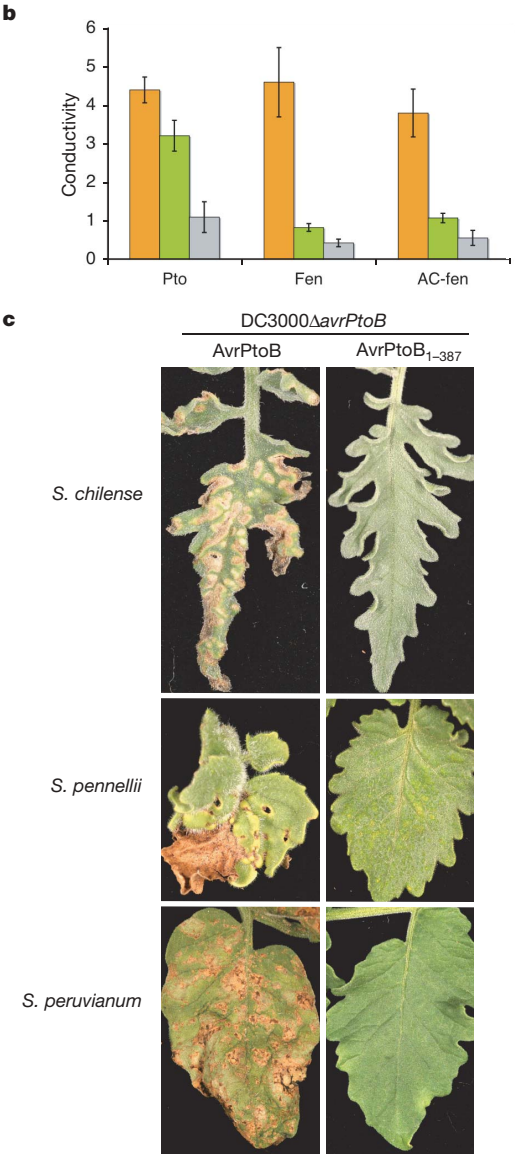


Figure 4 | Rsb is present in many cultivated and wild species of tomato. **a**, Inoculation of cultivated tomato varieties MM, VFNT-C or AC with *Pst* strains. R and S indicate resistant or susceptible plants, respectively. **b**, Electrolyte leakage after coexpression of AC-fen, Fen or Pto with AvrPtoB$_{1–387}$ (orange), AvrPtoB$_{1–307}$ (green) or AvrPtoB (grey). The ratio of conductivity of samples coexpressing AvrPtoB$_{1–307}$, AvrPtoB$_{1–387}$ or AvrPtoB versus empty vector control is shown. Error bars represent s.e.m. ($n = 3$). **c**, Leaves from 3 of 21 wild tomato species shown to exhibit Rsb immunity in response to inoculation with *Pst* delivering AvrPtoB$_{1–387}$ (see Supplementary Table 1): left, disease; right, Rsb immunity.

Our results can be viewed in the context of the evolutionary processes that may have shaped this pathogen–host interaction (Supplementary Fig. 7). It is known that the N-terminal region of AvrPtoB is able to suppress host basal defences[17,18]. This region promotes pathogen virulence and therefore an N-terminal-only form of AvrPtoB might have existed independently of the C-terminal E3 ligase domain. We now know that Fen binds the AvrPtoB$_{1–387}$ region to negate basal defence suppression, activate defence signalling and confer immunity on the host. Immunity mediated by Fen and by Pto requires Prf, suggesting that Prf evolved to function with a progenitor of the Pto family. It is possible this progenitor, unlike Fen, had a function in basal defence. An alternative is that members of the Pto family evolved together with Prf solely for effector-triggered immunity. To counter recognition by Fen, disrupt immunity-associated PCD and restore basal defence suppression activity, the N-terminal region of AvrPtoB may have acquired an E3 ligase domain to mediate the degradation of Fen and Fen orthologues. A similar ability to disrupt immunity has been reported for AvrRpt2, which disrupts signalling by the resistance protein RPM1 in *Arabidopsis* by cleaving an RPM1-interacting protein, RIN4 (ref. 26). It is possible that some AvrPtoB truncated proteins evolved to evade Fen recognition. For example, *P. syringae* pv. *maculicola* expresses an AvrPtoB homologue (HopPmaL[27]) that lacks both the E3 ligase domain and part of the Fen recognition determinant (residues 307–387). HopPmaL elicits immunity in response to Pto but not Fen[28]. Pto seems to have evolved to counter both of these strategies. It remains to be determined how Pto evades AvrPtoB-mediated ubiquitination.

METHODS SUMMARY

Plant inoculations. Tomato plants were vacuum infiltrated with *Pst* (about 5.5×10^4 colony-forming units per ml) suspended in 10 mM $MgCl_2$ and 0.00002% Silwet. Bacterial leaf populations were measured from three plants per treatment, three days after infiltration.

Virus-induced gene silencing (VIGS) and *Agrobacterium*-mediated transient expression. VIGS was induced by using the tobacco rattle virus vector delivered by *Agrobacterium tumefaciens*[14]. For transient gene expression, *A. tumefaciens* was used to deliver a 35S cauliflower mosaic virus expression cassette (pTEX)[4]. Ion leakage was measured two days after infiltration with *A. tumefaciens*.

Yeast two-hybrid assay. A LexA-based two-hybrid system was used to test for interactions between Pto family members in the bait vector and AvrPtoB or AvrPtoB mutants or truncations in the prey vector.

***In vitro* ubiquitination assay.** The *in vitro* ubiquitination reactions were performed with recombinant maltose-binding protein (MPB)-tagged Pto family proteins, His_6-tagged E1, His_6-tagged E2, ubiquitin and glutathione *S*-transferase (GST)-tagged AvrPtoB.

Protoplast bioassays. Protoplasts were isolated from seedling leaves and transformed with pTEX by using polyethylene glycol. Protoplast viability was determined by staining with Evans Blue.

Immunoblotting quantification. After detection of immunolabelled proteins by chemiluminescence, signal intensities were quantified by the Storm blot imaging system and quantified using ImageQuant TL software.

Plant material. Accessions of wild species of tomato were obtained from the Tomato Genetics Resource Center (http://tgrc.ucdavis.edu/).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 April; accepted 29 May 2007.

1. Mudgett, M. B. New insights to the function of phytopathogenic bacterial type III effectors in plants. *Annu. Rev. Plant Biol.* **56,** 509–531 (2005).
2. Abramovitch, R. B., Anderson, J. C. & Martin, G. B. Bacterial elicitation and evasion of plant innate immunity. *Nature Rev. Mol. Cell Biol.* **7,** 601–611 (2006).
3. Janjusevic, R., Abramovitch, R. B., Martin, G. B. & Stebbins, C. E. A bacterial inhibitor of host programmed cell death defenses is an E3 ubiquitin ligase. *Science* **311,** 222–226 (2006).
4. Abramovitch, R. B., Kim, Y.-J., Chen, S., Dickman, M. B. & Martin, G. B. *Pseudomonas* type III effector AvrPtoB induces plant disease susceptibility by inhibition of host programmed cell death. *EMBO J.* **22,** 60–69 (2003).
5. Buell, C. R. *et al.* The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Proc. Natl Acad. Sci. USA* **100,** 10181–10186 (2003).

6. Pedley, K. F. & Martin, G. B. Molecular basis of Pto-mediated resistance to bacterial speck disease. *Annu. Rev. Phytopathol.* **41,** 215–243 (2003).
7. Martin, G. B. *et al.* Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* **262,** 1432–1436 (1993).
8. Mucyn, T. S. *et al.* The tomato NBARC-LRR protein Prf interacts with Pto kinase *in vivo* to regulate specific plant immunity. *Plant Cell* **18,** 2792–2806 (2006).
9. Martin, G. B. *et al.* A member of the *Pto* gene family confers sensitivity to fenthion resulting in rapid cell death. *Plant Cell* **6,** 1543–1552 (1994).
10. Kim, Y.-J., Lin, N.-C. & Martin, G. B. Two distinct *Pseudomonas* effector proteins interact with the Pto kinase and activate plant immunity. *Cell* **109,** 589–598 (2002).
11. Abramovitch, R. B., Janjusevic, R., Stebbins, C. E. & Martin, G. B. Type III effector AvrPtoB requires intrinsic E3 ubiquitin ligase activity to suppress plant cell death and immunity. *Proc. Natl Acad. Sci. USA* **103,** 2851–2856 (2006).
12. Chang, J. H. *et al.* Functional analysis of the *Pto* resistance gene family in tomato and the identification of a minor resistance determinant in a susceptible haplotype. *Mol. Plant Microbe Interact.* **15,** 281–291 (2002).
13. Pascuzzi, P. E. *Structure-based functional analyses of Pseudomonas type III effector protein AvrPto and evaluation of putative virulence targets in tomato.* PhD thesis, Cornell Univ. (2006).
14. del Pozo, O., Pedley, K. F. & Martin, G. B. MAPKKKα is a positive regulator of cell death associated with both plant immunity and disease. *EMBO J.* **23,** 3072–3082 (2004).
15. Jia, Y., Loh, Y.-T., Zhou, J. & Martin, G. B. Alleles of Pto and Fen occur in bacterial speck-susceptible and fenthion-insensitive tomato and encode active protein kinases. *Plant Cell* **9,** 61–73 (1997).
16. Riely, B. K. & Martin, G. B. Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene *Pto*. *Proc. Natl Acad. Sci. USA* **98,** 2059–2064 (2001).
17. He, P. *et al.* Specific bacterial suppressors of MAMP signaling upstream of MAPKKK in *Arabidopsis* innate immunity. *Cell* **125,** 563–575 (2006).
18. de Torres, M. *et al. Pseudomonas syringae* effector AvrPtoB suppresses basal defence in *Arabidopsis. Plant J.* **47,** 368–382 (2006).
19. Badel, J. L., Shimizu, R., Oh, H. S. & Collmer, A. A *Pseudomonas syringae* pv. *tomato* avrE1/hopM1 mutant is severely reduced in growth and lesion formation in tomato. *Mol. Plant Microbe Interact.* **19,** 99–111 (2006).
20. Kim, M. G. *et al.* Two *Pseudomonas syringae* type III effectors inhibit RIN4-regulated basal defense in *Arabidopsis. Cell* **121,** 749–759 (2005).
21. Stebbins, C. E. & Galan, J. E. Structural mimicry in bacterial virulence. *Nature* **412,** 701–705 (2001).
22. Angot, A., Vergunst, A., Genin, S. & Peeters, N. Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems. *PLoS Pathogens* **3,** e3 (2007).
23. Rytkonen, A. & Holden, D. W. Bacterial interference of ubiquitination and deubiquitination. *Cell Host Microbe* **1,** 13–22 (2007).
24. Rohde, J. R., Breitkreutz, A., Chenal, A., Sansonetti, P. J. & Parsot, C. Type III secretion effectors of the IpaH family are E3 ubiquitin ligases. *Cell Host Microbe* **1,** 77–83 (2007).
25. Jamir, Y. *et al.* Identification of *Pseudomonas syringae* type III effectors that can suppress programmed cell death in plants and yeast. *Plant J.* **37,** 554–565 (2004).
26. Kim, H. S. *et al.* The *Pseudomonas syringae* effector AvrRpt2 cleaves its C-terminally acylated target, RIN4, from *Arabidopsis* membranes to block RPM1 activation. *Proc. Natl Acad. Sci. USA* **102,** 6496–6501 (2005).
27. Guttman, D. S. *et al.* A functional screen for the type III (Hrp) secretome of the plant pathogen *Pseudomonas syringae*. *Science* **295,** 1722–1726 (2002).
28. Lin, N. C., Abramovitch, R. B., Kim, Y. J. & Martin, G. B. Diverse AvrPtoB homologs from several *Pseudomonas syringae* pathovars elicit Pto-dependent resistance and have similar virulence activities. *Appl. Environ. Microbiol.* **72,** 702–712 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Munkvold and X. Tang for critical review of the manuscript; B. Randall for assistance with plant inoculations; J. Cohn and P. Pascuzzi for generation and characterization of the RG-PtoR(hpPto) line; J. Li and X. Tang for unpublished data; S. Collier for technical assistance; and our greenhouse staff for plant care. This work was supported, in part, by the NIH, the NSF and the Triad Foundation (G.B.M.).

Author Contributions T.R.R. and G.B.M. conceived, designed, and analysed the experiments. T.R.R. performed all of the experiments with the exceptions noted below. L.Z. performed the experiment shown in Figure 2b and Supplementary Fig. 5b. J.J.B. performed the experiments shown in Figure 1c and Supplementary Fig. 2. R.B.A. and F.X. provided technical assistance and unpublished clones. T.R.R. and G.B.M. wrote the manuscript. All authors contributed comments that were incorporated into the final version.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.B.M. (gbm7@cornell.edu).

METHODS

Virus-induced gene silencing. Gene silencing was induced with a tobacco rattle virus (TRV) vector delivered by *Agrobacterium tumefaciens* strain GV3101, as described previously[14]. *Pto* and *Pto* family member genes were silenced by using a conserved nucleotide sequence corresponding to *Pto* nucleotides 444–793. TRV2-silencing constructs of *Prf*, *MAPKKKa*, *MEK1*, *MEK2*, *SIPK* and *WIPK* have been described previously[14]. *A. tumefaciens* ($D_{600} = 0.15$) was induced with acetosyringone and used to infiltrate two-week-old *N. benthamiana* seedlings. Plants were maintained for five weeks at 18 °C with a 16-h daylength to allow silencing to occur.

Agrobacterium-mediated transient expression. *A. tumefaciens* strain GV2260 was used to deliver the pBTEX 35S cauliflower mosaic virus promoter expression cassette for transient gene expression. Pto, Fen and Ac-fen contain a C-terminal HA epitope tag. The presence or absence of PCD was apparent three days after infiltration into *N. benthamiana*. For cell death assays in *N. benthamiana*, three 1.2-cm diameter leaf discs per plant (three plants per experiment) were collected from infiltrated areas 48 or 72 h after infiltration. PCD was observed with co-expression of Pto and $AvrPtoB_{1–307}$ three days after infiltration; however, data were taken two days after infiltration with $AvrPtoB_{1–387}$ to avoid PCD caused by endogenous *N. benthamiana* Rsb signalling, which occurred three days after infiltration. Leaf discs were incubated in water at 23 °C with gentle shaking, and conductivity was measured after 4 h with an Acorn Con 5m (Oakton Instruments) as described previously[14]. The means of the ratios are shown and represent three experimental plants. Error bars indicate s.e.m. Experiments were repeated three times.

Yeast two-hybrid assays. A LexA-based two-hybrid system was used to test for interactions between Pto family members cloned into the bait vector, pEG202, and AvrPtoB or AvrPtoB mutants/truncations, Quad (K512R/K520R/K521R/K529R) or AvrPtoB-E2BS (E2BS, F479A/F525A/P533A), cloned into the prey vector, pJG4-5. Yeast cells were cotransformed by using the small-scale LiAC transformation procedure (Clontech). Transformants were selected on plates with leucine (Leu+). Successful transformants were grown on Leu+, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal), galactose medium plates for two days at 30 °C. Total protein was extracted with overnight cultures normalized to a D_{600} of 0.5 before being lysed. Experiments were repeated a minimum of three times.

Pathogenesis assays in tomato. Before inoculation, all *Pseudomonas syringae* pv. *tomato* strains were grown for two days on King's B (KB) medium plates containing rifampicin (100 mg l^{-1}) and any additional selection (DC3000Δ*avrPto* and DC3000:$AvrPtoB_{1–507}$, spectinomycin 50 μg ml^{-1}; DC3000Δ*avrPtoB* plus pCPP45; DC3000Δ*avrPtoB* plus pCPP45:$AvrPtoB_{1–387}$; and DC3000Δ*avrPtoB* plus pCPP45:$AvrPtoB_{1–307}$, kanamycin 50 μg ml^{-1}, tetracycline 10 μg ml^{-1}). Each bacterial strain was resusupended in 1 ml of liquid KB medium and standardized to a D_{600} of 0.1. Of this resuspension, 100 μl was spread onto fresh KB selection plates and allowed to grow for 18 h. Bacteria were collected and resuspended in 10 mM $MgCl_2$, and cultures were adjusted to a D_{600} of 0.00005 (about 5.5×10^4 colony-forming units per ml) in 10 mM $MgCl_2$, 0.00002% Silwet. Tomato plants were transplanted into 5.5-inch pots two weeks after seed sowing and one to two weeks before infiltration. Vacuum infiltrations and measurement of bacterial populations were performed as described previously[29]. Each suspension was serially diluted and colony-forming units were counted to verify equal bacterial inoculum levels. Bacterial populations in leaves were assessed with six leaf samples taken from each of three experimental plants per treatment three days after infiltration. All experiments were repeated at least three times. Experiments with DC3000Δ*hrcQ-U* (ref. 19) were the same as described above except that final cultures were adjusted to a D_{600} of 0.5 (about 5.5×10^8 colony-forming units per ml) in 10 mM $MgCl_2$, 0.025% Silwet, and plants were dipped in this bacterial suspension for 30 s instead of being vacuum infiltrated. Accessions of wild species of tomato were obtained from the Tomato Genetics Resource Center (http://tgrc.ucdavis.edu/). Accession numbers for the leaves shown in Fig. 4c are LA1960 (*S. chilense*), LA1272 (*S. pennellii*) and LA0446 (*S. peruvianum*).

Protoplast assays. Protoplasts were isolated from leaves of four-week-old Rio Grande-prf3 (RG-prf3), RG-PtoS[4] or *N. benthamiana*, and polyethylene glycol-mediated transformation was performed as described previously (ref. 30 and http://genetics.mgh.harvard.edu/sheenweb/). In brief, each gene was expressed in protoplasts by using the pTEX cauliflower mosaic virus 35S promoter expression cassette. *Pto*, *Fen* and *PtoC* expression was achieved with 20 μg of vector DNA, with an equal quantity of *AvrPtoB* or *AvrPtoB-Quad* vector DNA. However, 25 μg of vector DNA was used for *Fen* and *AvrPtoB* or *AvrPtoB-Quad* coexpression in experiments shown in Fig. 3. Protoplast treatments were as follows: DMSO (represents no treatment), MG132 (Sigma) was added at 50 μM, and plant protease inhibitor cocktail was added at 0.2% of the final volume (catalogue no. P9599; Sigma). All treatments contained a final concentration of 0.2% DMSO. Treatments occurred 20 h after transformation and progressed for 4 h. Protoplasts were lysed 24 h after transformation in lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM EDTA pH 8.0, 1.0% Triton X-100, 10% glycerol and 0.2% protease inhibitor). All experiments presented represent a minimum of three independent replicates.

Protoplast viability assay. To assess protoplast cell death, aqueous Evans Blue stain (0.04%) was added to RG-PtoS or *N. benthamiana* protoplasts 18 or 36 h, respectively, after transformation. Non-viable protoplasts were stained blue within 5 min and were detected by microscopy, counted, and recorded as a percentage of the total cells present.

Immunoblotting. To confirm protein expression in *N. benthamiana* tissue, in protoplasts and in yeast, total protein was resolved by 10% SDS–PAGE and transferred to poly(vinylidene difluoride) membrane (0.45 μm pore size; Millipore). Protein input was equalized for samples extracted from *N. benthamiana* (20 μg) and for lysed protoplasts (12.5 μg). HA-tagged Fen, Pto and PtoC were detected with anti-HA primary antibody (Roche Molecular Biochemicals); AvrPtoB or Quad, $AvrPtoB_{1–387}$ and $AvrPtoB_{1–307}$ were detected with a polyclonal anti-AvrPtoB primary antibody[11]; Pto family bait (pEG202) fusion proteins were detected with an anti-LexA primary antibody (Invitrogen); and ubiquitin was detected with a polyclonal anti-ubiquitin antibody (Santa Cruz Biotechnology). This was followed by chemiluminescence visualization (ECL kit; GE Healthcare Bio-Sciences Corp.). Signal intensities were detected with the Storm blot imaging system and quantified with ImageQuant TL software (GE Healthcare Bio-Sciences Corp.). Coomassie blue staining (0.2%) was used to verify equal loading in plant assays by assessing the abundance of Rubisco.

***In vitro* ubiquitination assay.** The *in vitro* ubiquitination assays were performed as described[31,32], with some modifications. In brief, the MBP-, GST-, or His_6-tagged proteins of interest were expressed in BL21 (DE3) *Escherichia coli* cells (Stratagene) and affinity-purified with an amylose (New England Biolabs), glutathione (Amersham, now GE Healthcare Bio-Sciences Corp.) or HIS-Select Nickel (Sigma-Aldrich) matrix, respectively. The *in vitro* ubiquitination reactions were performed by adding 200 ng of substrate protein (any of MBP-Pto, MBP-Fen, MBP-PtoC, MBP-PtoD, MBP-PDK1, MBP-AC-fen or MBP alone), 20 ng of purified His_6-E1 (wheat E1, GI: 136632), 100 ng of purified His_6-E2 (*Arabidopsis* ubiquitin-conjugating enzyme 8, AtUBC8), 12 μg of ubiquitin (BioMol International L.P.), and 0.6 μg of purified GST–AvrPtoB in the ubiquitination buffer (0.05 M Tris-HCl pH 7.5, 5 mM ATP, 5 mM $MgCl_2$, 2 mM dithiothreitol, 3 mM creatine phosphate, 5 μg ml^{-1} creatine phosphokinase) to a final volume of 30 μl. The reactions were incubated at 30 °C for 1.5 h before being stopped with SDS sample loading buffer and heated to 100 °C for 5 min. Half-volumes of the reactions were then separated by 10% SDS–PAGE. Ubiquitinated substrates were detected by western blotting with anti-MBP monoclonal antibody (New England Biolabs) followed by detection by chemiluminescence with the ECL kit. The experiment presented was repeated three times with similar results.

Mutagenesis and construct development. Amino acid substitutions in Fen-MBP were introduced using Pfu turbo polymerase PCR based site directed mutagenesis with primer pairs containing the following codon changes (in bold).

Fen K70R, 5′-AAGGTCGCCCTGAAA**AGG**CATAAACCTGAGTCC-3′ (for Fen K72R, 5′-GCCCTGAAAAAGCAT**AGA**CCTGAGTCCTCACAAG-3′ (forward) and 5′-CTTGTGAGGACTCAGG**TCT**ATGCTTTTTCAGGGC-3′ (reverse); Fen K155R, 5′-CTTCACTACCTTCAT**AGG**AATGCAGTTATACAT-3′ (forward) and 5′-ATGTATAACTGCATT**CCT**ATGAAGGTAGTGAAG-3′ (reverse); Fen K253R, 5′-GATGATGAGACGCAG**AGG**ATGGGACAGTTGGAA-3′ (forward) and 5′-TTCCAACTGTCCCAT**CCT**CTGCGTCTCATCATC-3′ (reverse); Fen K290R, 5′-TTAGCTCCGTCTAGT**AGA**AATAGGCCATCAATG-3′ (forward) and 5′-CATTGATGGCCTATT**TCT**ACTAGACGGAGCTAA-3′ (reverse).

29. Anderson, J. C., Pascuzzi, P. E., Xiao, F., Sessa, G. & Martin, G. B. Host-mediated phosphorylation of type III effector AvrPto promotes *Pseudomonas* virulence and avirulence in tomato. *Plant Cell* **18**, 502–514 (2006).
30. Xing, T., Malik, K., Martin, T. & Miki, B. L. Activation of tomato PR and wound-related genes by a mutagenized tomato MAP kinase kinase through divergent pathways. *Plant Mol. Biol.* **46**, 109–120 (2001).
31. Leng, R. P. *et al.* Pirh2, a p53-induced ubiquitin-protein ligase, promotes p53 degradation. *Cell* **112**, 779–791 (2003).
32. Zeng, L. R. *et al. Spotted leaf11*, a negative regulator of plant cell death and defense, encodes a U-box/armadillo repeat protein endowed with E3 ubiquitin ligase activity. *Plant Cell* **16**, 2795–2808 (2004).

LETTERS

Delayed ageing through damage protection by the Arf/p53 pathway

Ander Matheu[1]*†, Antonio Maraver[1]*, Peter Klatt[1], Ignacio Flores[2], Isabel Garcia-Cao[1], Consuelo Borras[3]†, Juana M. Flores[4], Jose Viña[3], Maria A. Blasco[2] & Manuel Serrano[1]

The tumour-suppressor pathway formed by the alternative reading frame protein of the *Cdkn2a* locus (Arf) and by p53 (also called Trp53) plays a central part in the detection and elimination of cellular damage, and this constitutes the basis of its potent cancer protection activity[1,2]. Similar to cancer, ageing also results from the accumulation of damage and, therefore, we have reasoned that Arf/p53 could have anti-ageing activity by alleviating the load of age-associated damage. Here we show that genetically manipulated mice with increased, but otherwise normally regulated, levels of Arf and p53 present strong cancer resistance and have decreased levels of ageing-associated damage. These observations extend the protective role of Arf/p53 to ageing, revealing a previously unknown anti-ageing mechanism and providing a rationale for the co-evolution of cancer resistance and longevity.

The tumour suppressor p53 is a master transcriptional factor that integrates and responds to a multitude of stresses[1]. In the case of severe stress, as occurs during cancer, p53, together with its positive regulator Arf, activates a terminal response that eliminates damaged cells from the proliferative pool by inducing either cellular senescence or apoptosis[2]. The type of damage associated with the physiological process of ageing is of low intensity, but its chronic nature eventually leads to a global decline in cellular functionality and tissue performance. Given the central role of Arf/p53 in sensing stress, it is conceivable that p53 could also sense ageing-associated damage (Supplementary Fig. 1). Recently, it has been shown in the nematode worm *Caenorhabditis elegans* that diverse mutations that increase longevity also increase cancer resistance through activation of p53 (ref. 3). Here we explore in mice the reciprocal concept, namely whether increasing the potency of the Arf/p53 tumour-suppression module translates into delayed ageing. Understanding the interplay between ageing and cancer is of importance for the development of therapeutic interventions against them.

We have previously shown that bacterial artificial chromosome-based genomic DNA transgenesis of a single extra gene-dose of *p53* or *Arf* protects mice against cancer[4,5]; we refer to these mice as super-p53 (s-p53) and super-Arf (s-Arf), respectively. In the case of s-Arf mice, the extra gene dose of *Arf* is accompanied by an extra gene dose of two other tumour suppressor genes, namely *Ink4a* and *Ink4b*, encoding the cyclin-dependent kinase inhibitors $p16^{Ink4a}$ and $p15^{Ink4b}$, respectively; however, as will be argued below, the anti-ageing phenotypes reported here are probably caused by Arf and, for simplicity, we will refer to these mice as s-Arf. These single-copy transgenes, s-*p53* and s-*Arf*, are expressed and regulated in a manner very similar to their endogenous counterparts and, therefore, result in moderately higher levels of the corresponding proteins throughout their lifespan, including at old age (see refs 4–6, and Supplementary Fig. 2a). It is well established that Arf stabilizes p53 and, on the basis of this, we wondered whether the combined effects of both transgenes (s-*p53* and s-*Arf*) in s-Arf/p53 mice (generated by crossing s-Arf and s-p53 mice) would result in a further increase in cancer protection compared to the previously reported cancer-resistance phenotypes of s-p53 and s-Arf mice[4,5]. First, we observed that the basal levels of p53 and p21, used as a molecular readout of the activity of p53, were elevated in s-Arf/p53 mice compared to s-Arf or s-p53 mice (Supplementary Fig. 2b, c). After this, we analysed the susceptibility of the cells derived from these mice to immortalization and neoplastic transformation *in vitro*, both of which are processes suppressed by the Arf/p53 module[4,5].

Primary mouse embryo fibroblasts (MEFs) derived from s-Arf/p53 mice showed a marked resistance to spontaneous immortalization (Fig. 1a). Interestingly, spontaneous immortalization always occurred in association with loss of the Arf/p53 module and in a manner that reflected the stochastic chance of losing *Arf* or *p53*. Thus, immortal wild-type MEFs lost *Arf* or *p53* with similar frequencies; s-Arf MEFs preferentially lost *p53*; s-p53 MEFs preferentially lost *Arf*; and, finally, s-Arf/p53 MEFs, like wild-type MEFs, lost *Arf* or *p53* with similar frequencies (Supplementary Table 2), although immortalization occurred after many passages and only in a fraction of cultures (Fig. 1a). Also, s-Arf/p53 MEFs were completely refractory to neoplastic transformation by the combined action of the adenoviral oncoprotein E1a and oncogenic Ras (Fig. 1b). More importantly, this increased resistance to immortalization and neoplastic transformation at the cellular level was paralleled by significantly diminished incidence of sporadic cancer in aged mice (Fig. 1c, and see Supplementary Fig. 3 for the spectra of tumour types). When compared with s-p53 or s-Arf mice, s-Arf/p53 mice showed a delay in the latency of chemically induced fibrosarcomas (Fig. 1d) and papillomas (Fig. 1e), and a decrease in the total number of papillomas (Fig. 1e). Together, these data demonstrate that increasing the gene dosage of *Ink4b-Arf-Ink4a* and *p53* results in a significant enhancement in cancer resistance, which, in the case of ageing-associated sporadic cancers, is strongly synergistic.

Our previous analyses of the lifespan of s-Arf and s-p53 mice indicated that, individually, the corresponding transgenes do not have a negative impact on normal ageing and lifespan[4–6]. Similarly, other investigators have reported that mice with low levels of Mdm2, and hence increased levels of p53, do not have a negative impact on longevity[7]. It is important to emphasize that these mouse models beyond having higher levels of p53 maintain intact the regulatory mechanisms that control p53 stability and activity. In contrast, other

[1]Tumor Suppression Group, [2]Telomeres and Telomerase Group, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain. [3]Department of Physiology, University of Valencia, Valencia 46010, Spain. [4]Department of Animal Surgery and Medicine, Complutense University of Madrid, Madrid 28040, Spain. †Present address: Division of Stem Cell and Developmental Genetics, MRC National Institute for Medical Research, Mill Hill, London NW7 1AA, UK (A.M.); Catholic University of Valencia, 94 Guillem de Castro Street, Valencia 46003, Spain (C.B.).
*These authors contributed equally to this work.

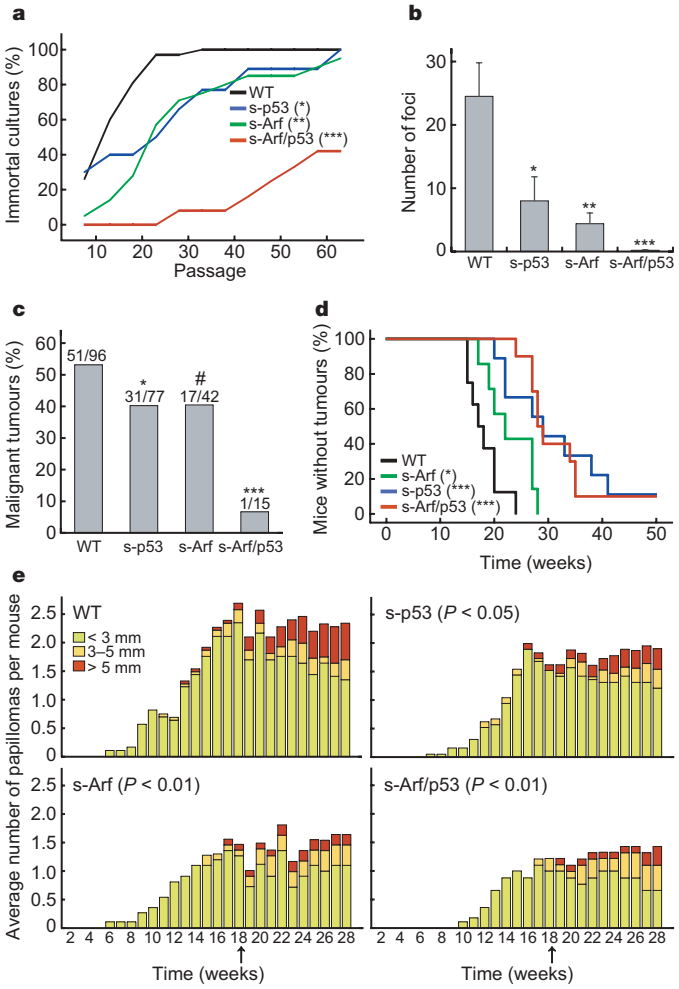


Figure 1 | The tumour suppressors Arf and p53 cooperate in conferring cancer resistance. **a**, The figure represents the percentage of cultures that were immortal at the indicated passage. Independent cultures of primary MEFs from wild-type (WT, $n = 38$), s-p53 ($n = 10$), s-Arf ($n = 21$) or s-Arf/p53 ($n = 10$) embryos were serially cultivated according to the 3T3 protocol. Cultures entered crisis and eventually resumed growth, which was scored as the point of immortalization. To assess statistical significance, we used Wilcoxon–Mann–Whitney rank sum test relative to wild type. See Supplementary Table 2 for the status of Arf and p53 in the spontaneously immortalized cultures. **b**, Independent cultures of primary MEFs from wild-type ($n = 6$), s-p53 ($n = 4$), s-Arf ($n = 5$) or s-Arf/p53 ($n = 13$) embryos were retrovirally transduced with H-RasV12 and E1a. Foci were scored after three weeks by Giemsa staining. Data are mean values ± s.e.m.; Student's *t*-test is relative to wild type. **c**, Wild-type ($n = 96$), s-p53 ($n = 77$), s-Arf ($n = 42$) or s-Arf/p53 ($n = 15$) mice were euthanized when they showed overt signs of poor health, such as reduced activity or dramatic weight loss, and were analysed for the presence of malignancies. Data are given as the percentage of mice that presented malignant tumours (see also Supplementary Fig. 3). We used Fisher's exact test relative to wild type. **d**, Wild-type ($n = 8$), s-p53 ($n = 9$), s-Arf ($n = 7$) or s-Arf/p53 ($n = 10$) mice were treated with 3-methylcholanthrene (3MC) and followed up for a period of 50 weeks until the appearance of fibrosarcomas. Kaplan–Meier representation of the survival curves of the four groups. We used logrank test relative to wild type. **e**, Papilloma formation in wild-type ($n = 17$), s-p53 ($n = 19$), s-Arf ($n = 11$) or s-Arf/p53 ($n = 9$) mice was initiated by a single application of 7,12-dimethylbenzanthracene (DMBA) when mice were 1–2 months old and, two weeks later, was followed by promotion with 12-*O*-tetradecanoylphorbol-13-acetate (TPA) for 16 weeks (treatment was discontinued at week 18 and is indicated with an arrow). The number and size of skin papillomas was scored weekly. Wilcoxon–Mann–Whitney rank sum test for each group is relative to wild type. Statistical significance: $\#P < 0.1$; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

mouse models expressing truncated versions of p53 that lack the Mdm2-binding site seem to have constitutive p53 activity and undergo premature ageing[8,9].

Given the substantial cooperation between Arf and p53 in cancer protection (see above), we wondered whether s-Arf/p53 mice could present alterations in ageing. For this, we scored the survival of cohorts of mice of the four relevant genotypes, all housed in the same animal facility and with the same genetic background, namely C57BL/6. Mice that were s-p53 had identical survival to wild-type mice, whereas s-Arf mice showed a modest increase in survival (Supplementary Fig. 4). The longevity of s-Arf/p53 mice was significantly extended (Fig. 2a), with an increase in median lifespan of 16% (see Supplementary Table 1). This effect can be put into context by comparison with the effect of calorie restriction on median lifespan observed in the C57BL/6 strain (20–25%; data obtained from refs 10 and 11), or by attenuation of growth hormone signalling in C57BL/6 mice (9–20%; data obtained from ref. 12). An important difference between the above-mentioned mouse models and s-Arf/p53 mice is that maximum lifespan was not increased in s-Arf/p53 mice, indicating that the Arf/p53 pathway alleviates some, but not all, of the causes of physiological ageing. To discount the possible contribution of cancer to the lifespan of the different mouse colonies, we considered separately the lifespan of cancer-free mice. Importantly, cancer-free s-Arf/p53 mice retained extended longevity compared to cancer-free wild-type mice (Fig. 2b, Supplementary Fig. 4, and Supplementary Table 1). Moreover, the spectrum of ageing-associated diseases, including benign tumours, in s-Arf/p53 mice was indistinguishable from that in wild-type mice (Supplementary Fig. 5), albeit at a more advanced age (Supplementary Fig. 6). These observations suggest that the Arf/p53 module has an anti-ageing effect, although it cannot be excluded that the anti-tumoral activity of Arf/p53 could also contribute to the extended average lifespan of s-Arf/p53 mice.

To further substantiate the delayed ageing of s-Arf/p53 mice, we analysed a series of biomarkers of ageing in mice of the four relevant genotypes and at different ages. Neuromuscular coordination declines with ageing[13]; this decline was significantly attenuated in s-Arf/p53 mice, whereas s-p53 and s-Arf mice presented an intermediate performance (Fig. 2c). Similarly, s-Arf/p53 mice had a higher capacity to re-grow hair, even at very old ages (Fig. 2d). It has recently been shown, both in mice and in primates, that ageing is associated with the presence of phosphorylated histone H2AX (γ-H2AX) foci in cells, which are thought to reflect sites of DNA lesions[14,15]. Using this molecular marker of ageing, we observed that s-Arf/p53 mice have decreased age-dependent accumulation of γ-H2AX-positive cells in liver and kidney, whereas s-p53 and s-Arf mice present an intermediate accumulation of this marker (Fig. 2e and Supplementary Fig. 7). Together, these data further support that the concomitant increase of Arf and p53 delays ageing. With regard to *Ink4a*, which is also present in the s-*Arf* transgene, current evidence indicates that *Ink4a* limits the long-term proliferative capacity of various adult stem cells in aged mice[16–18]. Therefore, if any, the impact of *Ink4a* on ageing should be pro-ageing, and this would cause an underestimation of the anti-ageing effect of Arf/p53.

We have explored several possible mechanisms that could underlie the observed delay in ageing of s-Arf/p53 mice, namely decreased metabolism, decreased insulin-like growth factor 1 (IGF-1) levels, decreased telomere shortening and decreased oxidative damage. Regarding the first two mechanisms, we did not detect changes in s-Arf/p53 mice compared to wild-type controls with respect to body weight, food and water intake, output of faeces and urine, and the levels of serum IGF-1 (Supplementary Fig. 8a–c). We detected a slight increase in telomere length in the liver of old s-Arf/p53 mice compared to wild-type controls, although the difference did not reach statistical significance (Supplementary Fig. 8d). Importantly, we found decreased generation of reactive oxygen species (ROS) in splenocytes of s-Arf/p53 mice, both at young and old ages (Fig. 3a and

Supplementary Fig. 9). We hypothesized that decreased generation of ROS in s-Arf/p53 mice should translate into decreased accumulation of oxidative damage. Indeed, we observed that age-dependent accumulation of oxidized lipids in the liver was diminished in s-Arf/p53 mice (Fig. 3b and Supplementary Fig. 10). Similarly, the levels of oxidized proteins were also lower in s-Arf/p53 mice at young and old ages, with an intermediate level in the case of s-p53 and s-Arf mice (Fig. 3c and Supplementary Fig. 11). Together, these findings support the notion that the Arf/p53 pathway confers protection against ageing-associated oxidative damage.

Mice expressing catalase in the mitochondria constitute the most compelling demonstration of the anti-ageing activity of antioxidant defences in mammals[19]. Recently, it has been reported that p53 activates an antioxidant transcriptional programme that could be

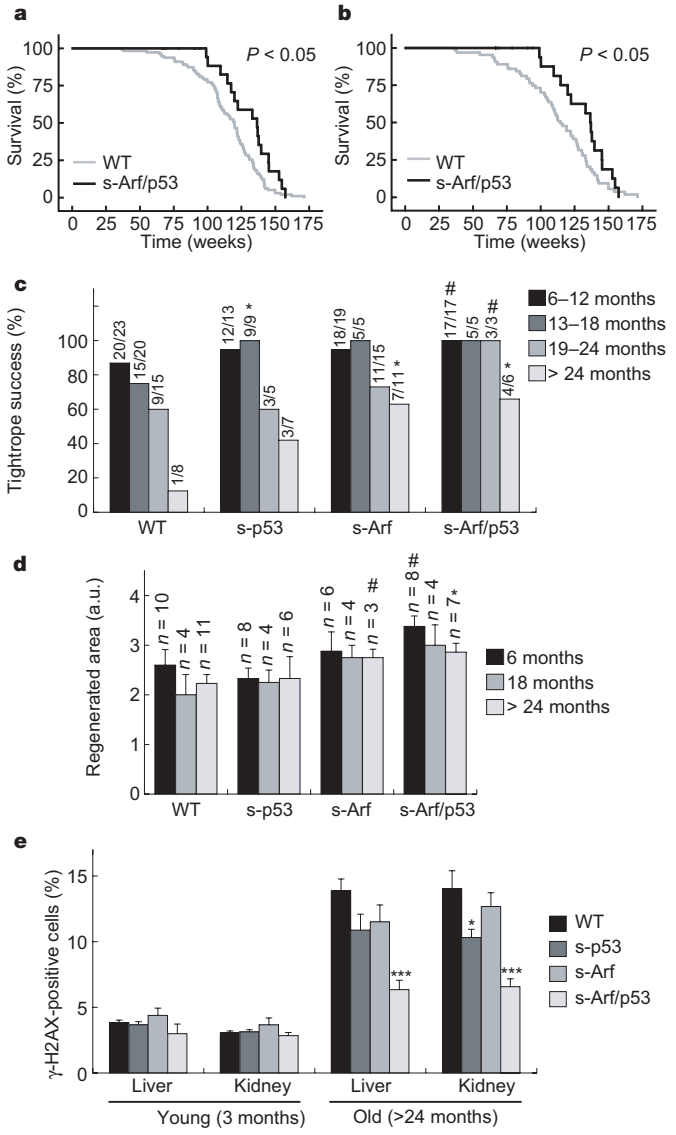


Figure 2 | Delayed ageing in s-Arf/p53 mice. **a**, Survival curves of cohorts of wild-type ($n = 111$) and s-Arf/p53 ($n = 25$) mice. Shown is the Kaplan–Meier representation of the two groups. **b**, The survival curves of the two mouse cohorts from **a** were redrawn, excluding those animals that presented malignant tumours at the time of death. For wild type, $n = 64$, and for s-Arf/p53, $n = 24$. Statistical significance was assessed using the logrank test. For additional data see Supplementary Table 1 and Supplementary Fig. 4. **c**, Neuromuscular coordination was quantified as the percentage of mice that successfully passed the tightrope test (see Supplementary Information). The fraction of mice passing the test is indicated above the bars. Fisher's Exact test for each age group is relative to wild type. **d**, Hair re-growth capacity of the dorsal skin was quantified in arbitrary units (a.u., see Supplementary Information) 15 d after plucking. Fisher's Exact test for each age group is relative to wild type. **e**, DNA damage. Liver and kidney cryosections from young ($n = 2$ for each genotype) and aged ($n = 3$–5 for each genotype) mice were analysed for the percentage of γ-H2AX-positive nuclei by confocal microscopy. Data are mean values ± s.e.m.; Student's *t*-test is relative to wild type. Statistical significance: $\#P < 0.1$; $*P < 0.05$; $***P < 0.001$.

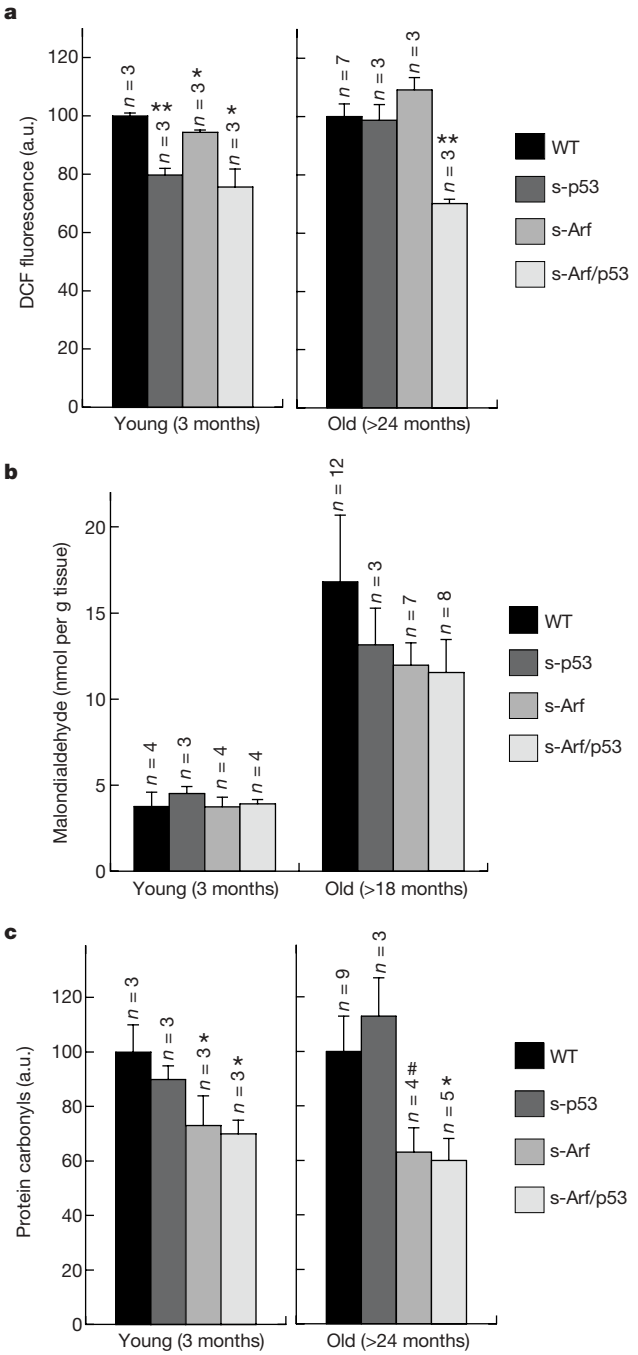


Figure 3 | Decreased oxidative damage in s-Arf/p53 mice. **a**, Splenocytes were analysed for ROS levels by FACS analysis using DCF. Data are mean values ± s.e.m., and normalized for each age group to ROS levels in wild-type splenocytes. Student's *t*-test is relative to wild type. **b**, Lipid peroxidation was measured in liver. Liver samples were analysed for the presence of malondialdehyde (MDA) using HPLC. **c**, Levels of oxidized proteins were measured in the liver. Liver samples were analysed for the presence of protein carbonyls by immunoblotting. Data are mean values ± s.e.m.; Student's *t*-test for each age group is relative to wild type. Statistical significance: $\#P < 0.1$; $*P < 0.05$; $**P < 0.01$.

particularly relevant when p53 is activated by mild stresses[20], such as those presumed to operate during physiological ageing. Of prominence in the p53 antioxidant programme are the antioxidant genes sestrin (*Sesn*)*1* and *Sesn2* (refs 21–23). This family of proteins has a fundamental role in maintaining the activity of peroxiredoxins, which in turn are critical antioxidant defences[24–26]. In confirmation and extension of the above, we found that *Sesn1* and *Sesn2* are expressed in a p53-dependent manner in MEFs, reaching the highest levels in s-Arf/p53 MEFs (Fig. 4a). *Arf*-only-null MEFs had lower levels of sestrins compared to wild-type MEFs (Fig. 4a), suggesting that Arf, rather than Ink4a or Ink4b, is responsible for the positive effect of the s-*Arf* transgene on the expression of sestrins. Similarly, expression of sestrins was moderately upregulated in the liver of both young and old s-Arf/p53 mice (Fig. 4b). We could not detect significant changes in the levels of other genes with anti-ageing activity, such as *Shc1* (also known as $p66^{Shc}$) or *Sirt1* (Supplementary Fig. 12). The expression of p53 targets, including sestrins and *p21*, was not significantly affected by ageing (Supplementary Fig. 13). We interpret this as meaning that, despite the accumulation of damaged macromolecules and cells with ageing (see above for γ-H2AX, lipids and proteins), the rate of damage generation and signalling to p53 is age-independent, and the protective activity of p53 is constant throughout the lifespan. The antioxidant effect of Arf/p53 was further supported by the observation that the levels of reduced glutathione (GSH), which is a measure of the global antioxidant activity, were increased in young and old s-Arf/p53 mice, both in liver (Fig. 4c and Supplementary Fig. 14) and in brain (Supplementary Fig. 15). Finally, we performed a functional assay *in vivo* to evaluate the anti-oxidant activity of Arf/p53. The survival time of mice after acute oxidative stress reflects the potency of the antioxidant defences and is known to correlate with longevity[27–29]. On the basis of this, young wild-type and s-Arf/p53 mice were injected with a lethal dose of paraquat, a strong oxidative agent, and their survival time was scored (Fig. 4d). The results obtained indicate that Arf/p53 mice have increased resistance to oxidative damage, providing additional *in vivo* evidence of the anti-oxidant activity of Arf/p53.

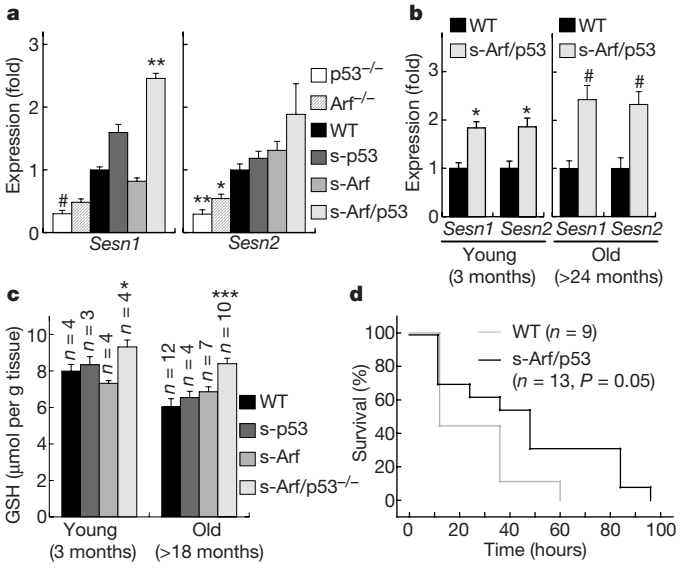


Figure 4 | Increased expression of antioxidant genes in s-Arf/p53 mice. **a**, Expression of *Sesn1* and *Sesn2* was measured in MEFs grown in 3% O_2 by quantitative real-time PCR (see Supplementary Information). PCR determinations were performed with MEFs from 3–5 independent MEF cultures per genotype. Values are expressed relative to wild type. Student's *t*-test is relative to wild type. **b**, Expression of *Sesn1* and *Sesn2* in liver samples ($n = 4$–7 for each age and genotype) was determined by quantitative real-time PCR. Student's *t*-test for each age group is relative to wild type. **c**, GSH levels were determined by HPLC in liver samples. Data are mean values ± s.e.m.; Student's *t*-test for each age group is relative to wild type. **d**, Mice were intraperitoneally injected with a lethal dose of paraquat (60 mg kg^{-1}), and their survival fraction was scored every 12 h. Statistical significance was assessed using the logrank test. Statistical significance: #$P < 0.1$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Other investigators have found that permanent activation of p53 results in premature ageing[8,9], and that the absence of p53 may alleviate the premature ageing of mice with high levels of constitutive endogenous damage[30]. It is of critical importance to note that, in these mouse models of premature ageing, p53 is permanently activated, owing either to truncation of p53 domains or to constitutive endogenous damage. When p53 activity is enhanced but normal regulation is retained there is no acceleration of ageing, as testified by three independent mouse models with increased p53 (ref. 4), increased Arf (ref. 5) or decreased Mdm2 (ref. 7). More tellingly, as we report here, a combined increase in Arf and p53 results in detectable anti-ageing activity. We propose that the spectra of genes activated by p53 under normal physiological conditions have a global anti-oxidant effect, thus decreasing ageing-associated oxidative damage. Our observations, together with those recently reported for *C. elegans*, demonstrate that p53 intimately links longevity and cancer resistance. Thus, mutations that extend the lifespan of *C. elegans* result in increased activity of p53 and cancer resistance[3]; and, conversely, increasing the activity of p53 in mice produces both cancer resistance and delayed ageing (in this report). The simultaneous impact of p53 on cancer resistance and longevity could explain, at least in part, the co-evolution of these two features across species.

METHODS SUMMARY

Handling, supervision and experimentation with mice was done in accordance to the Guidelines for Humane Endpoints for Animals Used in Biomedical Research. Assays with MEFs were carried out according to standard practice. Carcinogenic assays were performed as previously described[4,5]. Protein and messenger RNA levels were measured by conventional immunoblotting and quantitative real-time polymerase chain reaction (PCR). γ-H2AX was measured by immunofluorescence using cryosections[6]. Lipid peroxidation and glutathione were measured by HPLC. ROS levels were quantified as 5-(and-6-)-chloromethyl-2′,7′-dichlorodihydrofluorescein diacetate acetyl ester (DCF) fluorescence using a FACS. Oxidized proteins were measured by immunoblotting using an antibody that recognizes carbonylated proteins.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 March; accepted 22 May 2007.

1. Harris, S. L. & Levine, A. J. The p53 pathway: positive and negative feedback loops. *Oncogene* **24,** 2899–2908 (2005).
2. Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432,** 307–315 (2004).
3. Pinkston, J. M., Garigan, D., Hansen, M. & Kenyon, C. Mutations that increase the life span of *C. elegans* inhibit tumor growth. *Science* **313,** 971–975 (2006).
4. Garcia-Cao, I. *et al.* 'Super p53' mice exhibit enhanced DNA damage response, are tumor resistant and age normally. *EMBO J.* **21,** 6225–6235 (2002).
5. Matheu, A. *et al.* Increased gene dosage of *Ink4a/Arf* results in cancer resistance and normal aging. *Genes Dev.* **18,** 2736–2746 (2004).
6. Garcia-Cao, I. *et al.* Increased p53 activity does not accelerate telomere-driven ageing. *EMBO Rep.* **7,** 546–552 (2006).
7. Mendrysa, S. M. *et al.* Tumor suppression and normal aging in mice with constitutively high p53 activity. *Genes Dev.* **20,** 16–21 (2006).
8. Tyner, S. D. *et al.* p53 mutant mice that display early ageing-associated phenotypes. *Nature* **415,** 45–53 (2002).
9. Maier, B. *et al.* Modulation of mammalian life span by the short isoform of p53. *Genes Dev.* **18,** 306–319 (2004).
10. Forster, M. J., Morris, P. & Sohal, R. S. Genotype and age influence the effect of caloric intake on mortality in mice. *FASEB J.* **17,** 690–692 (2003).
11. Liang, H. *et al.* Genetic mouse models of extended lifespan. *Exp. Gerontol.* **38,** 1353–1364 (2003).
12. Coschigano, K. T. *et al.* Deletion, but not antagonism, of the mouse growth hormone receptor results in severely decreased body weights, insulin, and insulin-like growth factor I levels and increased life span. *Endocrinology* **144,** 3799–3810 (2003).
13. Ingram, D. K. & Reynolds, M. A. Assessing the predictive validity of psychomotor tests as measures of biological age in mice. *Exp. Aging Res.* **12,** 155–162 (1986).

14. Sedelnikova, O. A. *et al.* Senescing human cells and ageing mice accumulate DNA lesions with unrepairable double-strand breaks. *Nature Cell Biol.* **6,** 168–170 (2004).
15. Herbig, U., Ferreira, M., Condel, L., Carey, D. & Sedivy, J. M. Cellular senescence in aging primates. *Science* **311,** 1257 (2006).
16. Janzen, V. *et al.* Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16INK4a. *Nature* **443,** 421–426 (2006).
17. Krishnamurthy, J. *et al.* p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature* **443,** 453–457 (2006).
18. Molofsky, A. V. *et al.* Increasing p16INK4a expression decreases forebrain progenitors and neurogenesis during ageing. *Nature* **443,** 448–452 (2006).
19. Schriner, S. E. *et al.* Extension of murine life span by overexpression of catalase targeted to mitochondria. *Science* **308,** 1909–1911 (2005).
20. Vousden, K. H. & Lane, D. P. p53 in health and disease. *Nature Rev. Mol. Cell Biol.* **8,** 275–283 (2007).
21. Velasco-Miguel, S. *et al.* PA26, a novel target of the p53 tumor suppressor and member of the GADD family of DNA damage and growth arrest inducible genes. *Oncogene* **18,** 127–137 (1999).
22. Budanov, A. V. *et al.* Identification of a novel stress-responsive gene *Hi95* involved in regulation of cell viability. *Oncogene* **21,** 6017–6031 (2002).
23. Sablina, A. A. *et al.* The antioxidant function of the p53 tumor suppressor. *Nature Med.* **11,** 1306–1313 (2005).
24. Neumann, C. A. *et al.* Essential role for the peroxiredoxin Prdx1 in erythrocyte antioxidant defence and tumour suppression. *Nature* **424,** 561–565 (2003).
25. Wood, Z. A., Schroder, E., Robin Harris, J. & Poole, L. B. Structure, mechanism and regulation of peroxiredoxins. *Trends Biochem. Sci.* **28,** 32–40 (2003).
26. Budanov, A. V., Sablina, A. A., Feinstein, E., Koonin, E. V. & Chumakov, P. M. Regeneration of peroxiredoxins by p53-regulated sestrins, homologs of bacterial AhpD. *Science* **304,** 596–600 (2004).
27. Agarwal, S. & Sohal, R. S. Relationship between susceptibility to protein oxidation, aging, and maximum life span potential of different species. *Exp. Gerontol.* **31,** 365–372 (1996).
28. Kapahi, P., Boulton, M. E. & Kirkwood, T. B. Positive correlation between mammalian life span and cellular resistance to stress. *Free Radic. Biol. Med.* **26,** 495–500 (1999).
29. Yamamoto, M. *et al.* Regulation of oxidative stress by the anti-aging hormone klotho. *J. Biol. Chem.* **280,** 38029–38034 (2005).
30. Varela, I. *et al.* Accelerated ageing in mice deficient in Zmpste24 protease is linked to p53 signalling activation. *Nature* **437,** 564–568 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Muñoz for mouse colony management and animal care and E. Santos for mouse genotyping. A. Matheu was funded by a predoctoral fellowship from the Spanish Ministry of Education and Science (MEC); A. Maraver is funded by the Juan de la Cierva Program of the MEC; and I.F. is funded by the Ramon y Cajal Program of the MEC. This work has been funded by the MEC (M.S. and M.A.B.) and the European Union (M.S. and M.A.B.). M.A.B. is a recipient of the Josef Steiner Cancer Research Award 2003.

Author Contributions A. Matheu and A. Maraver contributed equally to this work; A. Matheu, A. Maraver, I.F., I.G.-C., C.B. and J.M.F. performed experimental work; P.K. analysed data and assisted in editing the paper; M.S. wrote the paper; J.V. and M.A.B. co-directed research; and M.S. designed research and directed the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (mserrano@cnio.es).

METHODS

Mice and primary mouse cells. Mice were housed at the pathogen-free barrier area of the Spanish National Cancer Research Center, Madrid. Mice were observed on a daily basis and were euthanized when they showed overt signs of morbidity or tumours, in accordance with the Guidelines for Humane Endpoints for Animals Used in Biomedical Research. Double-transgenic s-Arf/p53 mice were generated by crossing single-transgenic s-p53 (ref. 4) and s-Arf (ref. 5) mice. The genetic background of all the mice used in this study is pure C57BL/6J, including the s-p53 and s-Arf mice (used to generate the s-Arf/p53), which had previously been backcrossed for seven generations with wild-type C57BL/6J mice.

The tightrope test is a widely used and extensively validated behavioural marker of ageing[13]. We performed the test with slight variations: mice were placed on a bar of circular section (60 cm long and 1.5 cm diameter) and the test was considered successful when a mouse did not fall during a period of 60 s in at least one out of five consecutive trials. For the hair re-growth assay, dorsal hair was removed by plucking from a square of approximately 1.5 cm × 1.5 cm. Hair re-growth was scored two weeks later on the basis of digital photographs and a semi-quantitative assessment using an arbitrary scale from one to four (where four represents complete hair regeneration). Scoring was performed blindly by two investigators, who obtained essentially identical scores.

MEFs were prepared from day 13.5 embryos and cultured as previously described[31,32]. For Supplementary Table 2, the status of Arf and of p53 was determined by immunoblot before and after treatment of the cells with doxorubicin as previously described[31]. Neoplastic transformation of MEFs was assessed by focus formation assays of MEFs retrovirally transduced with E1a and HRasV12 as previously described[5].

Chemical carcinogenesis and paraquat tolerance test. For 3MC-induced carcinogenesis, four-month-old mice received a single intramuscular injection into one of their rear legs of a 40 μl solution containing 3MC (Sigma), dissolved at a final concentration of 25 μg μl^{-1} in sesame oil (Sigma) as previously described[4,5,33]. Mice were observed on a daily basis, and were euthanized when tumours reached a diameter of at least 1.5 cm. For DMBA/TPA-induced carcinogenesis, mice of 1–2 months received a single application of 25 μg DMBA (Sigma). Two weeks later, mice were treated twice per week during 16 additional weeks with 5 μg TPA (Sigma) dissolved in acetone. For the paraquat tolerance test, we basically followed a previously reported procedure[29]. In particular, 5-month-old mice were injected intraperitoneally with 60 mg kg^{-1} of paraquat (methylviologendichloride hydrate, Sigma) diluted in PBS. Survival was scored every 12 h.

Protein analyses. Cell extracts were prepared by incubation of cells on ice for 10 min in a 50 mM Tris-HCl buffer (pH 7.5), supplemented with 150 mM NaCl, NP-40 0.5%, 1 mM phenylmethylsulphonyl fluoride, 2 mg ml^{-1} aprotinin, 2 mg ml^{-1} leupeptin and 1 mg ml^{-1} pepstatin followed by removal of cellular debris by centrifugation at 13,400*g* for 10 min. Protein concentration was measured using BioRad DC Protein Assay Kit. For immunoblots, samples corresponding to 30 μg protein were resolved on 4–20% SDS–PAGE gels, wet-transferred to nitrocellulose (BioRad) and immunoblotted. For immunoprecipitation, 1 mg total protein was subjected to immunoprecipitation with 2.4 μg anti-p53 (PAb 246, Santa Cruz Biotechnology). The following antibodies were used for immunoblotting: for detection of Arf, R562 or ab10569 from AbCam; for p53, CM-5 from Novocastra; for p21, C-19 from Santa Cruz; and, for β-actin, AC-15 from Sigma. Secondary antibodies were either horseradish peroxidase-linked anti-rabbit (DAKO) or anti-mouse (DAKO). Detection was performed by chemiluminescence using the ECL detection system (Amersham).

Quantitative real-time RT–PCR. Total RNA from tissues was extracted with Trizol (Life Technologies). Samples were treated with DNAseI before reverse transcription using random priming and Superscript Reverse Transcriptase (Life Technologies), according to the manufacturer's guidelines. Quantitative real-time PCR was performed using an ABI PRISM 7700 (Applied Biosystems), using DNA Master SYBR Green I mix (Applied Biosystems). The primers used were: Sesn1-F, 5′-GTCTGGATAACATCACATTAG-3′; Sesn1-R, 5'-CCAGGTAGGAACACTGATGC-3′. Sesn2-F, 5′-CTCACAGCTGGTCTGTGTG-3′; Sesn2-R, 5′-CCTCCGTGTGGCAATACC-3′. p16-F, 5′-AACTCTTTCGGTCGTACCCC-3′; p16-R, 5′-GCGTGCTTGAGCTGAAGCTA-3′. p15-F, 5′-GTCATGATGATGGGCAGCG-3′; p15-R, 5′-GCGTGCAGATACCTCGC-3′. Arf-F, 5′-GCCGCACCGGAATCCT-3′; Arf-R, 5′-TTGAGCAGAAGAGCTGCTACGT-3′. p21-F, 5′-GTGGGTCTGACTCCAGCCC-3′; p21-R, 5′-CCTTCTCGTGAGACGCTTAC-3′. Sirt1-F, 5′-GCTGACGACTTCGACGACG-3′; Sirt1-R, 5′-TCGGTCAACAGGAGGTTGTCT-3′. p66-F, 5′-GGCTGCATCCCAACGACAA-3′; p66-R, 5′-CCTGAGTCCGGGTAT TGAAGT-3′. Act-F, 5′-GGCACCACACCTTCTACAATG-3′; Act-R, 5′-GTGGTGGTGAAGCTGTAGCC-3′.

The difference in PCR cycles with respect to β-actin (ΔCt) for a given experimental sample was subtracted from the corresponding ΔCt of the reference sample (such as wild-type) (ΔΔCt). Statistical analyses (Student's t-test) were performed on the ΔΔCt as recommended by previous investigators[35]. For representation purposes, ΔΔCt were converted into fold expression ($2^{\Delta\Delta Ct}$). The error bars correspond to the relative error of the ΔΔCt values.

Measurement of oxidative damage, DNA damage, telomere length and IGF-1 levels. Intracellular levels of ROS were determined by flow cytometry using the ROS-specific fluorescent probe DCF (Invitrogen). In summary, cells were incubated for 20 min at 37 °C with DCF at a final concentration of 10 μg ml^{-1} in PBS containing 5 mM glucose. After washing the cells twice with PBS and incubation in culture medium for an additional 20 min at 37 °C, fluorescence intensity was measured using a FACScalibur (excitation 488 nm, emission 515–545 nm). Data were analysed with CELLQuest software. GSH levels were determined by HPLC as previously described[36]. Lipid peroxidation was determined as accumulation of MDA, which was detected by HPLC as an MDA–thiobarbituric acid adduct[37]. Oxidative modification of total proteins was assessed by immunoblot detection of protein carbonyl groups using the 'OxyBlot' protein oxidation kit (Intergen) following the manufacturer's instructions. The procedure to quantify total protein carbonyls with the OxyBlot kit was densitometry of the oxyblot and of the Ponceau staining, followed by finding the ratio between the total density in the oxyblot and the total density in the Ponceau. DNA damage was assessed by confocal immunofluorescence against γ-H2AX (antibody clone JBW301 from Upstate Biotechnology) on cryosections as previously described[6]. Telomere length in hepatocytes was determined by quantitative fluorescence *in situ* hybridization analysis of interphase nuclei in cryosections from mouse livers essentially as described previously for paraffin-embedded skin[38]. Serum levels of IGF-1 were measured using retro-orbital blood and the Rat IGF-1 kit from Diagnostic Systems Laboratories.

31. Pantoja, C. & Serrano, M. Murine fibroblasts lacking p21 undergo senescence and are resistant to transformation by oncogenic Ras. *Oncogene* **18,** 4974–4982 (1999).
32. Palmero, I. & Serrano, M. Induction of senescence by oncogenic Ras. *Methods Enzymol.* **333,** 247–256 (2001).
33. Wexler, H. & Rosenberg, S. A. Pulmonary metastases from autochthonous 3-methylcholanthrene-induced murine tumors. *J. Natl Cancer Inst.* **63,** 1393–1395 (1979).
34. Matheu, A., Klatt, P. & Serrano, M. Regulation of the *INK4a/ARF* locus by histone deacetylase inhibitors. *J. Biol. Chem.* **280,** 42433–42441 (2005).
35. Yuan, J. S., Reed, A., Chen, F. & Stewart, C. N. Jr Statistical analysis of real-time PCR data. *BMC Bioinformatics* **7,** 85 (2006).
36. Asensi, M., Sastre, J., Pallardo, F. V., Estrela, J. M. & Vina, J. Determination of oxidized glutathione in blood: high-performance liquid chromatography. *Methods Enzymol.* **234,** 367–371 (1994).
37. Wong, S. H. *et al.* Lipoperoxides in plasma as measured by liquid-chromatographic separation of malondialdehyde-thiobarbituric acid adduct. *Clin. Chem.* **33,** 214–220 (1987).
38. Gonzalez-Suarez, E., Samper, E., Flores, J. M. & Blasco, M. A. Telomerase-deficient mice with short telomeres are resistant to skin tumorigenesis. *Nature Genet.* **26,** 114–117 (2000).

Modern mating

Match point.

John Zakour

Finally it was about to happen. All those hours in the lab, even more time invested searching, almost begging for grant money and backers, was about to pay off. A hundred years ago they had matchmakers. At the turn of century they had match-making websites. HA! Those were the dark ages! We were set to exponentially one-up them all with the very first custom made Perfect Android Mate, PAM-I.

Looking at her laying there on the lab table, she really was a sweet sight. Blonde hair draped over her shoulders, blue eyes with long curly lashes, perfectly chiselled button nose, a mouth with lips that were both pouty and sweet, and a body that had more curves than a road up the Alps. Every man would love to have her as his companion. And that's based on her looks alone.

But PAM-I is so much more than her appearance. Being programmable and customizable, we can make her whatever the end-user wants her to be. Her likes and interests can complement the user's perfectly. She gives new meaning to the term 'user-friendly'.

PAM-I was only the beginning! Once we had her up and running there would be other models of both sexes to follow. Yes sir, we were going to eliminate incompatibility and loneliness forever. I wasn't sure if I'd get a Nobel for this, but I was going to make a lot of cash.

Deb, my longest-serving postdoc, walked into the lab. Deb was a pleasant enough looking girl: short dark hair, friendly eyes; but she was no PAM-I. Then again, no flesh-and-blood woman was; they all came with their own likes and dislikes, their own mental baggage weighing them down. We were about to put an end to all that.

Deb looked PAM-I over from head to toe with a wary eye. "Dr Sebastian, are you sure about this?"

"Of course I'm sure. I have tenure!" I said.

She shook her head. "Man-made mates seem so unnatural." Deb considered herself my conscience.

Walking up and down the table bearing PAM-I, I gave her the last check-over. "This is 2077, we can do better than natural. The customer expects it."

Deb shook her head. "If you say so, I'm just a lowly postdoc."

"You've entered the last small-talk and useful trivia subroutines?" I asked.

Deb pulled out her small, paper-thin computer. She looked at it. "Yes."

I stroked PAM-I gently on the cheek; her skin was so soft. "You've alerted the press?"

She nodded. "Yes. Reps from all the major networks and blogs will be here in four hours." She hesitated. "Are you sure you want to show her to the press so soon? We haven't even done any beta testing yet."

I smiled at her. "I'll be doing the beta testing myself. After all, she is programmed to be my perfect match."

Deb lowered her head. "Are you really that forlorn?" she asked meekly.

JACEY

I shrugged. It's true science can be a lonely mistress, especially when you're not the best-looking guy around. Humans can be so superficial. I couldn't admit that to one of my employees though. "No, no, of course not. Many scientists use themselves in experiments. Jung and Freud did it all the time. It's a prudent cost-cutting measure."

Deb rolled her eyes but I ignored her. I pointed to PAM-I. "Time to activate her!" I said dramatically.

Deb pushed a button on her computer.

"You are recording this for history," I reminded Deb.

"Of course," she sighed.

PAM-I shot up into a sitting position. She opened her eyes. "Where am I?" she asked.

"You are in New New York University artificial human lab-12," I told her.

"Oh," PAM-I said looking around. "Nice place."

"We're subsidized by Wal-Mart," Deb told her.

"So what's your favourite colour?" I asked PAM-I.

"Blue," she said.

"How many children would you like?" I asked.

"I can have children?"

"Yes, you are fully functional," Deb told her.

PAM-I smiled. "One girl, with red hair."

"Your favourite food?"

"Grilled shrimp with a touch of lime," she said.

"Who won the 1986 World Series?" I asked her.

"The Mets."

PAM-I was great. She was working out exactly as planned.

"What do you consider to be the perfect evening?" Deb asked.

PAM-I looked up at the ceiling thinking about her answer. She bit her lip and said: "Watching a ball game while giving my man a back rub and snacking on pizza."

Deb turned to me. "Yep, she's perfect for you."

"I know," I said barely able to contain my glee.

I held my hand out to PAM-I. She accepted it. I helped her to her feet. She took a step and stumbled. I caught her. She straightened up. "Sorry," she said. "Legs take a while to get used to."

"No problem," I said. I looked at my wrist computer. "We have three hours until I introduce you to the press."

"As your date?" she asked.

"Yes, that's the whole purpose of this operation," I said.

"To find you a date?" PAM-I asked, eyebrow raised.

"No, to make it so that everyone can have a perfect companion. You and I are the test pilots."

PAM-I looked at me with eyes wide open. She shook her head. "I don't like that."

Deb looked at me. "See, it's not right building an intelligent being for the sole purpose of being somebody's mate."

PAM-I shook her head again. "No, that's not what I mind at all…"

"Then what is it?" I asked.

"You're not nearly good looking enough for me," PAM-I told me.

Damn! We made her too human… ■

John Zakour is a SF/humour writer with a master's degree in human behaviour. His last novel was *The Frost-Haired Vixen* (Daw, 2006). His next two novels will be *The Blue-Haired Bombshell* (Daw, 2007) and *Baxter Moon Galactic Scout* (Brown Barn, 2008). He can be found at www.johnzakour.com.